Tracing Microbes and their Genes through the Environment: Diversity, Ecology and Evolution

Dissertation
zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)
vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich
von

SAMUEL CHAFFRON

aus Frankreich

Promotionskomitee:

Prof. Dr. Andreas Wagner
Dr. Peer Bork
Prof. Dr. Jakob Pernthaler
Prof. Dr. Christian von Mering
(Vorsitz und Leitung der Dissertation)

Zürich, August 2010

ACKNOWLEDGMENTS

First of all, I would like to thank my PhD thesis advisor Prof. Dr. Christian von Mering. Christian was supervising my master thesis in EMBL Heidelberg and he really influenced my decision to pursue my cursus and to start a PhD in his new research group at the University of Zurich. Christian, thank you so much for your great mentoring and your passion for science that you can communicate so well, I learned a lot from you.

I also wish to thank all the fishes I have been swimming with in the aquarium of our institute, working with you was always great fun!

Thank you to all my friends around the globe and in Zurich who always support me. Spéciale dédicace à la French crew et à la German class, merci pour tous ces bons moments passés ensemble! Bien sûr, merci à toi, Paulina, pour tout.

Finalement, je souhaite aussi remercier de tout mon cœur ma famille et tout particulièrement ma sœur Céline, Youssou et Noah, et mes parents, Maryse et André. Merci pour votre amour et votre continuel soutien.

Kenavo!

ABSTRACT

Microbial life developed on Earth approximately 3.5 billion years ago and has since become one of the fundamental engines driving the cycles of energy and matter on our planet. Microorganisms have also been intimately intertwined with the evolution of macroscopic organisms, and were among the main actors in shaping life as we know it today. They can improve human well-being by complex interactions taking place in our intestinal tract and on various body surfaces, but they can also lead to life-threatening situations in case of severe infections. Despite their global significance, scientists are just beginning to understand the structure and function of natural microbial communities. Still largely uncharacterized are their ecological roles, and how community structures feed back onto microbial evolution.

Today, high-throughput genome sequencing and functional genomics methods are revolutionizing the field of environmental microbiology and are reshaping our view on microbial ecosystems. A grand challenge is the global integration and mining of information at various system levels of organization – at the level of a single cell but also at the community level. Ultimately, the fusion of microbial ecology and 'omics' technologies will help us to better understand and characterize the microorganisms and ecosystems in and around us.

During the last 3 years, I have focused my efforts on developing concepts and methods to process and analyze genomics and proteomics data from environmental microbial communities. My work was motivated by the emergence of new high-throughput technologies that can generate very large amounts of data, thus making computational biology approaches indispensable to analyze and mine the information in this field. Diversity assessment of natural microbial communities is greatly improved by deep-sequencing of 16S ribosomal ribonucleic acid (rRNA) gene amplicons recovered from the environment. We developed an *in silico* pipeline in order to process and analyze sequences sampled from different systems (e. g., the mouse gut, pitcher plants, ...) in collaboration with several research groups. Using this tool, we were able, for example, to characterize the commensal microbiota in a mouse model and to show that the colonization-success of an extrinsic bacterial species (pathogenic or beneficial) into an established gut ecosystem is facilitated by the pre-existing abundance of closely related bacteria. This finding could help to increase the efficacy of probiotic therapy, and also to identify patients at risk of developing chronic enteric infections.

In a second project, in order to help microbiologists to mine genomics and proteomics datasets from microbial populations living on plant leaves, we developed a new method to map and integrate this type of information using complete reference genomes. This approach allows to assign functional and metabolic capabilities to lineages present in the microbial community, and this enabled us to reveal the nature and strategies of bacteria living in this habitat.

As we demonstrated in our analysis of the gut microbiota, interactions among microbes in a given ecosystem can play an important role in shaping the global structure of a community. Although ecological preferences of microbes are difficult to evaluate in the environment, using environmental and genomic sequence data we were able to globally assess, for the first time, habitat and coexistence preferences for a significant number of microbial taxa. Clearly, the distribution of microbes over the globe is not random (i. e., stable habitat preferences exist) and specific coexistences between lineages can be detected. Interestingly, coexisting lineages are more functionally similar than expected at random, underlining the significant impact of lineage groupings on genome evolution.

Taken together, the results presented in this thesis give insights into the structure, ecology and metabolism of microorganisms in their natural habitat.

ZUSAMMENFASSUNG

Mikrobielles Leben entwickelte sich auf der Erde vor ungefähr 3.5 Milliarden Jahren und ist bis heute die treibende Kraft hinter den Energie- und Materiezyklen auf unserem Planeten. Mikroorganismen waren die ersten Lebewesen und haben auch die Entwicklung des mehrzelligen Lebens, wie wir es heute kennen, entscheidend mitgeformt. Sie können die menschliche Gesundheit durch komplexe Interaktionen, z.B. in unseren Eingeweiden, beeinflussen. Trotz ihrer allumfassenden Bedeutung beginnen wir erst langsam die Struktur und Funktion von natürlichen mikrobiellen Gemeinschaften zu verstehen. Ebenfalls weitgehend uncharakterisiert sind ihre ökologischen Vorlieben und wie diese die Evolution der Mikroorganismen beeinflussen.

Heutzutage revolutionieren Hoch-Durchsatz-Genomsequenzierung und Daten aus der funktionellen Genomik das Feld der Mikrobiologie und unsere Sicht auf mikrobielle Ökosysteme. Eine grosse Herausforderung ist die umfassende Integration und Erschliessung von Information auf den verschiedenen Systemebenen der Organisation, sowohl in der einzelnen Zelle als auch auf der Ebene der Lebensgemeinschaft. Schlussendlich wird die Vereinigung von mikrobieller Ökologie und den 'omics'-Technologien uns helfen, Mikroorganismen, ihre Ökosysteme und ihre Evolution besser zu verstehen.

Während der letzten drei Jahre konzentrierten wir unsere Anstrengungen darauf, Konzepte und Methoden zu entwickeln, um Genom- und Proteomdaten von natürlichen mikrobiellen Gemeinschaften zu verarbeiten und zu analysieren. Diese Arbeit war durch das Erscheinen neuer Hoch-Durchsatz-Technologien motiviert, die sehr grosse Datenmengen produzieren und deshalb die Datenanalyse durch Bioinformatik-Ansätze voraussetzen. Die Beurteilung der Diversität natürlicher mikrobieller Gemeinschaften wird durch 'Deep Sequencing' von 16S ribosomaler RNA (rRNA), die aus Proben entnommen wird, wesentlich verbessert. Wir haben in Zusammenarbeit mit mehreren Forschungsgruppen eine *in silico* Pipeline entwickelt, um Sequenzen aus verschiedenen Systemen (z.B. Mausdarm, Blattoberflächen) zu verarbeiten und analysieren. Mit diesem Werkzeug waren wir in der Lage, z.B. kommensale

Mikroorganismen in einem Mausmodell zu charakterisieren und zu zeigen, dass der Besiedlungserfolg einer extrinsischen Bakterienart (krankheitserregend oder nützlich) in ein bereits etabliertes Darmökosystem von der relativen Menge bereits vorhanderen, nahe verwandter Bakterien beeinflusst wird. Diese Feststellung könnte helfen, die Effizienz von probiotischen Therapien zu steigern, und ebenso, Patienten zu identifizieren, die ein erhöhtes Risiko haben, Magen-Darm- Infektionen zu entwickeln.

In einem zweiten Projekt haben wir eine neue Methode entwickelt, die Mikrobiologen hilft, Genom- und Proteomdaten von mikrobiellen Populationen auf Pflanzenblättern zu untersuchen, indem die Daten auf komplette Referenzgenome projiziert werden. Dieser Ansatz erlaubt es, einzelnen Gruppen innerhalb der mikrobiellen Gemeinschaft funktionelle und metabolische Fähigkeiten zuzuweisen, was es uns ermöglicht hat, die Struktur und die Strategien von kommensalen Bakterien in diesem Habitat zu erschliessen.

Wie wir in unserer Analyse von Darm-Mikroorganismen gezeigt haben, können Interaktionen von Mikroorganismen in einem gegebenen Ökosystem eine wichtige Rolle bei der Stabilisierung der Gesamtstruktur der Gemeinschaft spielen. Obwohl ökologische Präferenzen von Mikroorganismen in freier Natur sehr schwierig zu beurteilen sind, waren wir, mit Hilfe von Genomsequenzen und Habitatbeschreibungen, in der Lage, zum ersten Mal Habitat- und Koexistenzpräferenzen für eine bedeutende Zahl von mikrobiellen Taxa fest zu stellen. Die Verteilung von Mikroben auf der Erde ist eindeutig nicht zufällig (d.h. Habitatpräferenzen existieren) und spezifische Koexistenzen zwischen einzelnen Abstammungslinien können festgestellt werden. Interessanterweise sind koexistierende Gruppen funktionell ähnlicher als erwartet, was den signifikanten Einfluss des Zusammenlebens von Mikroben auf die Evolution ihres Genoms unterstreicht.

Alles in allem geben die Ergebnisse in dieser Doktorarbeit einen Einblick in die Struktur, Ökologie und den Metabolismus von Mikroorganismen in ihrem natürlichen Lebensraum.

PUBLICATIONS

Publications included in this thesis:

Paper I
Stecher, Robbiani, Walker, Westendorf, Barthel, Kremer, Chaffron, Macpherson, Buer, Parkhill, Dougan, von Mering, and Hardt[111]
***Salmonella enterica* serovar typhimurium exploits inflammation to compete with the intestinal microbiota.**
PLoS Biology, **5(10)**:2177–2189, October 2007.

Paper II
Chaffron and von Mering[17]
Termites in the woodwork.
Genome Biology, **8(11)**:229, 2007.

Paper III
Delmotte*, Knief*, Chaffron, Innerebner, Roschitzki, Schlapbach, von Mering, and Vorholt[22]
Community proteogenomics reveals insights into the physiology of phyllosphere bacteria.
Proceedings of the National Academy of Sciences of the United States of America, **106 (38)**:16428–6433, September 2009.

Paper IV
Stecher*, Chaffron*, Käppeli*, Hapfelmeier, Freedrich, Weber, Kirundi, Suar, McCoy, von Mering, Macpherson, and Hardt[112]
Like will to like: abundances of closely related species can predict susceptibility to intestinal colonization by pathogenic and commensal bacteria.
PLoS Pathogens, **6(1)**:e1000711, January 2010.

Paper V
Chaffron, Rehrauer, Pernthaler, and von Mering[18]
A global network of coexisting microbes from environmental and whole-genome sequence data.
Genome Research, June 2010.

CONTENTS

ACRONYMS

AOM anaerobic oxidation of methane

BAC bacterial artificial chromosome

BLAST basic local alignment search tool

DGGE denaturing gradient gel electrophoresis

DNA deoxyribonucleic acid

FISH fluorescent in situ hybridization

HGT horizontal gene transfer

IS insertion sequence

LC liquid chromatography

MDA multiple displacement amplification

MS mass spectrometry

ORF open reading frame

OTU operational taxonomic unit

PCR polymerase chain reaction

RNA ribonucleic acid

rRNA ribosomal ribonucleic acid

SNP single nucleotide polymorphism

2D-PAGE two-dimensional polyacrylamide gel electrophoresis

Part I

INTRODUCTION AND CONTEXT

1 INTRODUCTION

Microbes are the most abundant and diverse living organisms on Earth, distributed among the three primary domains of Life: archaea, bacteria, and eukarya[79]. The Earth's ecosystem is balanced to no small extent by microbial activities: they make up more than half of the world's biomass. Indeed, microbial communities can achieve an extremely wide spectrum of complex reactions and processes. They play a major role in nutrient recycling and thus in the global element cycles of our ecosystem, such as the carbon, phosphorous and nitrogen cycles. Therefore, it is of major importance to characterize and understand microbes in their environments. In addition, they also need to be monitored and tightly controlled during medical practice, and in many biotechnology applications.

Since the first observation of a microorganism by Antonie van Leeuwenhoek in 1675, we know that microbes can be found virtually in all parts of the biosphere. More than 7000 named microbial species are currently considered accepted, as defined by genotypic and phenotypic properties[1]. But microbial diversity has been shown to be actually considerably higher, since it is monitored, for more than two decades now, using molecular-phylogenetic approaches – without the requirement for cultivation – in a great variety of habitats. Carl Woese was the first to establish a molecular sequence-based phylogenetic "Tree of Life" by comparison of rRNA sequences[130,129]. Today, most microbial diversity surveys are based on the 16S rRNA sequences and the leading reference in bacterial and archaeal taxonomy, the *Bergey's Manual of Systematic Bacteriology*, has adopted a 16S rRNA framework for classification[54].

Today, the development of high-throughput techniques to determine and monitor sequences, transcripts, proteins and metabolite levels, *in situ*, will help us to better understand and characterize the diversity, physiology and ecology of microbial communities in their environment. The emerging field of environmental genomics aims to give access to the "unseen majority" of microorganisms in nature[127,106] without cultivating them in a laboratory. Using new sequencing technologies, genome sequences can now be directly recovered from the environment. Then, diversity and gene functions can be analyzed, and correlated with specific environmental characteristics. This approach has already permitted measurements of diversity levels *in situ*[120,5] and determination of gene families specific to or enriched in a given habitat[115,31]. Applied to low-complexity microbial ecosystems, it is even possible to recover near-complete genomes of naturally occurring bacterial species[117].
Environmental genomics is a powerful approach for characterizing microbial communities and holds even greater potential when combined with other complementary technologies such as transcriptomics, metabolomics or proteomics[124]. For example, it is possible to use mass spectrometry (MS)-based proteomic methods to evaluate gene expression and partitioning of metabolic functions within natural microbial communities[92].

Systems biology can be defined as the study of complex interactions between the components of biological systems in order to understand and describe the function and behavior of that system. From this perspective, (low-complexity) microbial communities can constitute a great model that can be built upon and constrained by genome sequencing and high-throughput functional genomics data. Integration of this type of information allows to functionally characterize microbes in their habitat and can help us to better understand metabolic functioning and evolutionary dynamics of microbial communities.

Microbial ecology is an exciting and growing field, in which researchers aim to study the relationships that microorganisms can have among each other and with their environments. Two aspects stand out in particular: the microbial productivity and diversity in response to climate change, and the increasing realization that complex microbial communities can influence human health. In this context, it is clear that systems (micro)biology should be considered at the ecosystem level[90,121].

The work presented in this thesis represents a step forward towards this goal, by tracing microbes and their genes in various environments and developing new methods to understand functional roles within microbial communities and to detect microbial ecological preferences *in silico*.

2 ENVIRONMENTAL MICROBIOLOGY

2.1 SEQUENCES AND DIVERSITY

The pioneering work of Carl Woese and George Fox on the 16S rRNA gene changed our view of microbial diversity and evolution by providing an objective phylogenetic framework in which to classify cellular life[130,129]. Before the discovery of the 16S rRNA gene as a potential marker for microbial identification, microbiologists were limited by the requirement to isolate microorganisms in pure cultures, in order to subsequently analyze them using various physiological and biochemical traits. Quantification of cells in environmental samples was performed via viable plate counts or "most-probable-number" techniques[95]. These methods are clearly not appropriate for diversity estimation in environmental samples because they tend to select for certain organisms. This problem was coined the "great plate count anomaly" by Staley and Konopka[109]: direct microscopic counts frequently exceeded viable cell counts, by several orders of magnitude.

The first attempts to characterize natural microbial communities using rRNA began more than two decades ago. 5S rRNA molecules were directly extracted from environmental samples, separated by electrophoresis, sequenced and compared to known sequences by multiple sequence alignment and phylogenetic tree inference[107,108]. This approach gave interesting insights but the information coded in the approximately 120-nucleotide of the 5S rRNA is relatively limited and the requirement for electrophoretic separation restricted the application of this method to low-complexity microbial communities. Consequently, a new approach was proposed and developed by Pace and colleagues to tackle these problems. The 16S rRNA molecule has an average length of 1500 nucleotides and contains sufficient information for reliable phylogenetic analyses. They therefore developed a procedure to extract, clone and sequence the 16S gene from environmental samples in order to perform comparative analyses of the retrieved sequences. They first applied this approach to characterize a marine picoplankton community[103] and were able to identify numerous unknown sequences related to Cyanobacteria and Gamma-proteobacteria. Another approach took advantage of the polymerase chain reaction (PCR) to selectively amplified 16S rRNA gene fragments from environmental samples of picoplankton from the Sargasso sea[45] and also detected the presence of microorganisms related to Cyanobacteria and Proteobacteria. This method allowed the identification of not only new bacterial species but also of new archaeal species[41]. These studies demonstrated for the first time that previously unknown sequences can be retrieved from the environment and revealed that sequences deposited in 16S rRNA databases (at the time, mostly from culture collection strains) do not represent the naturally occurring diversity of microorganisms.

The sequence characteristics of the 16S rRNA genes (which differ along their lengths in term of relative sequence conservation, including several hyper-variable regions)

also enabled the development of important techniques to characterize microbial communities *in situ*. It is possible to target discrete regions of the 16S gene for hybridization to group- and species-specific oligonucleotides probes. These probes, coupled to fluorescent dies, allow the direct observation and identification of single cells by fluorescence microscopy. This technique, commonly called fluorescent in situ hybridization (FISH), was pioneered by DeLong et al. [24] and Amann et al. [3]. These developments made whole-cell hybridization with rRNA-targeted probes a suitable tool for well-defined, phylogenetic, and environmental studies in microbiology.
Sequence variation in the 16S rRNA genes can also be exploited to directly determine the genetic diversity of complex microbial populations. The procedure is based on the separation of PCR-amplified 16S deoxyribonucleic acid (DNA) fragments by electrophoresis in polyacrylamide gel containing a linearly increasing gradient of denaturants. This method, called denaturing gradient gel electrophoresis (DGGE), was developed by Muyzer et al. [70] and allows DNA fragments of same length but different in sequence to be separated, therefore giving a global picture of the genetic diversity within an uncharacterized microbial community. Both FISH and DGGE techniques are still in use today and contributed significantly to our better understanding of natural microbial communities. For more than two decades, these techniques enabled culture-independent molecular surveys (through the cloning and sequencing of 16S genes directly from the environment) that are continuously revealing the extensive complexity and diversity of the microbial world.

Although it is increasingly recognized that the 16S rRNA gene may have insufficient resolution for the reliable binning of microbes into species[1], today, this gene remains the marker of choice to characterize microbial community diversity and structure, *in situ*. The size of dedicated rRNA sequence databases[19,85,27] is growing at an ever increasing pace, due to the emergence of new sequencing technologies that provide a higher throughput at a lower cost.

Several analytical tools have been developed to compare diversity and structure among microbial communities[101,50]. A common approach is the definition of operational taxonomic units (OTUs), which allows the description of a given community at various levels of phylogenetic resolution, by clustering the 16S rRNA gene at various degrees of similarity. Therefore, this technique can be extremely useful to compare microbial communities across various samples.
OTUs also help to circumvent the problem of the microbial species definition, which constitutes an intensively debated topic at the moment[32]. The controversy is mainly due to the immense diversity and micro-diversity of microbes on Earth, but also to the fact that microorganisms can exchange DNA horizontally, a process commonly called horizontal gene transfer (HGT). New concepts to better define microbial species (e. g., using whole-genome sequence information) have been proposed but none have been widely accepted[44]. Generally, it becomes clearer that the actual definition of a microbial species is outdated and needs to be replaced with a framework that includes the environmental and ecological context. Indeed, genetic, population, and ecological parameters must all be taken into the equation as they all significantly drive the evolution of microbes.

Since the first application of whole-genome shotgun sequencing to natural microbial populations[120], this latter problem can partially be circumvented. It is now possible to access the genomic repertoire of a given microbial community in any environment: this is the basis for a newly emerging field called *metagenomics* or *environmental genomics*.

2.2 GENOMICS *in situ*

The Sanger sequencing technology, which was developed in the 70's by Frederick Sanger and co-workers[100] enabled the rapid sequencing of a multitude of genes and notably the determination of the first 16S rRNA complete sequence, from *Escherichia coli*[14]. Today, the technique is still used, for example for the traditional PCR cloning and sequencing of full-length 16S rRNA sequences.
The field of genomics was established when the first DNA-based complete genome, that of the bacteriophage phi-X174, was sequenced in 1977[99] using the Sanger sequencing method. In 1995, the first complete genome sequence of a free-living organism, *Haemophilus influenzae*, was determined by an innovative approach developed by Craig Venter and colleagues[37]. Instead of sequencing clones (lambda or cosmid clones) derived from mapped restriction fragments, they sequenced and assembled (using computational methods) randomly generated fragments of DNA of 300 to 500 bp length from the whole chromosome. The method called 'shotgun sequencing' constitutes a rapid, accurate and less expensive strategy that can be applied to whole genomes. That same year, the second complete genome of a free-living organism was completed by the same team, using the same methodology; the field of comparative genomics was born[38]. The availability, for the first time, of two complete genomes allowed the authors to compare the gene content and their organization and to describe a minimal set of genes that are required to sustain life. Moreover, they could also identify the genes responsible for physiological differences between both bacteria.
Today, thousands of complete genomes from viruses, bacteria and eukaryotes (including *Homo sapiens*) are publicly available on-line. And the advent of new high-throughput sequencing technologies that provide, for equivalent costs, several orders of magnitude more sensitivity is continuously increasing and enriching this mass of information.

The genomic analyses of natural microbial communities started with the isolation of large DNA fragments using bacterial artificial chromosome (BAC) libraries[105]. Several groups were successful in cloning large chromosome fragments (>100kpb) from uncultivated microbial groups sampled in the ocean[9] and in the soil[94]. In one of the first studies applying this methodology[8], the shotgun sequence analysis of a BAC clone from a widespread marine planktonic bacterium (SAR 86 group) revealed the first type of bacterial rhodopsin[8]. The biochemical and biophysical analyses of this novel bacterial rhodopsin demonstrated its function as a light-driven proton pump, revealing a new type of phototrophy in the world's oceans[10] and setting up the path for the new field of environmental genomics. In the following years, major discoveries have been enabled using a metagenomic approach. A good example is the genomic

evidence that the anaerobic oxidation of methane (AOM) in marine sediments, which consumes annually a significant amount of methane, can be performed by a specific group of archaea encoding all the steps of a methanogenic pathway[49].

Today, new sequencing technologies, such as pyrosequencing, enable the direct and extremely fast recovery of DNA fragments from environments, hence greatly facilitating the genomic analyses of natural microbial populations. The pyrosequencing technology (i. e., detection of pyrophosphate release with an enzymatic cascade including luciferase and ending by the measurement of the emitted light) was first developed as a genotyping tool for single nucleotide polymorphism (SNP) detection[34]. The main innovations necessary to prepare this technology for high-throughput, *de novo* sequencing were the miniaturization of the pyrosequencing reaction and moving both the template preparation step and the pyrosequencing chemistry to the solid phase, allowing at the same time the sequencing to be performed in parallel. It resulted in a major advancement in the art of sequencing by synthesis, achieving an approximately 100-fold increase in throughput over the current Sanger sequencing technology[64]. This new breakthrough lead to the development of the 454 sequencing platform, which, together with other next-generation sequencing technologies, has enabled a real democratization of high-throughput sequencing[97]. As of today, the original publication of Margulies et al. [64] counts over 1300 citations, illustrating the remarkably broad impact of the applications of this new method. And indeed, over the last years, the 454 sequencing technique allowed the determination of the first non-Sanger sequence of an individual human in only 2 months[126] and also considerably helped characterizing the genomic repertoire of natural microbial communities, for example of human-associated microbes[72]. The technology enabled as well a higher throughput for ribonucleic acid (RNA)-centered meta-transcriptomic approaches[39,118]. Other high-throughput sequencing technologies are also available, such as Illumina sequencing, which was recently applied to extensively catalogue the human gut microbial gene repertoire[87].

Other strategies are being experimented in order to access the genome sequence of microorganisms in their environments. Instead of massively sequencing a given environment, methods are developed to sequence an entire genome from a single, uncultured cell. This is relevant due to the difficulty of assembling contigs into discrete genomes, and also because environmental sampling is always biased towards abundant species. A single molecule of DNA can be amplified using multiple displacement amplification (MDA)[20] and then sequenced[136]. Another method takes advantage of PCR and a microfluidics device in order to amplify and analyze multiple genes from single cells isolated from the environment[77].

Ultimately, researchers are developing technologies to sequence individual, single molecules of DNA, without the need for cloning, amplification or ligation during sample preparation[11]. As of today, various methods have been published and have promised improvements in throughput[46,51,33] to the extent that a complete human genome could be sequenced using a single-molecule sequencing method[86]. More recently, a protocol to mark and visualize specific groups of microbes (FISH) has been modified in order to apply the methodology *in situ* and to enable the isolation of specific groups of cells via flow cytometry[135]. All these methods should enable and

facilitate the systems analysis of complex microbial communities, from the ecological level down to the molecular level.

From the amino-acids perspective, the development of MS methods to measure proteins in complex samples is also undergoing its own revolution.

2.3 PROTEOMICS *in situ*

The field of proteomics can be described as a collection of various methodologies to systematically detect and measure proteins[2]. Modern proteomics approaches largely trace back to the invention of two major techniques in the last decades: two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) and MS. Today, MS-based proteomics is an indispensable tool for systems biology, enabling the cataloguing of proteins[15,104] and the verification and functional annotation of many predicted genes[28,104], the analysis of protein expression and localization, as well as the analysis of protein complexes and the discovery of protein-protein interactions.

In microbial proteomics, the analysis of complex mixtures of proteins is performed via both gel-based and gel-independent liquid chromatography (LC)-based separations, followed by a MS-based peptide identification[121]. Proteins are then identified by the comparison of the measured intact masses and fragmentation patterns of the peptides with predicted ones, which are generated *in silico* from genomic sequence data. Environmental proteomics applied to low-complexity microbial communities has already been shown to be an extremely helpful approach to evaluate gene expression and characterize metabolic functions in order to understand the physiology and strategies of dominant microbial species[92].

Using this approach to characterize more complex ecosystems is challenging, because protein identification relies on the available genomic information, which is always incomplete given the extremely high species diversity in certain environments (e. g., soil, ocean). Another problem are the highly variable abundance levels of proteins: low abundance proteins are more difficult to detect using MS. Complete detection of the majority of unique protein products in a sample may even be impossible[128]. Some of these problems have been addressed recently by the application of shotgun MS-based whole community proteomics coupled to metagenomics, for several highly complex samples[122,22].

The applications of genomics and proteomics on microbial populations have just started but have already revealed new capabilities for studying the physiology, ecology and evolution of microbes in their environments.

2.4 MICROBIAL ECOLOGY

We just begin to appreciate the importance of microbial interactions and their global influence on the broad spectrum of habitats colonized by microorganisms. Microbes are main actors shaping and regulating various systems such as Earth's ecosystems

or human health, and therefore are essential to sustain life on our planet. It is crucial to discover and understand the various interactions that can occur among them. A striking example to illustrate this importance is the global significance of photosynthetic marine Cyanobacteria. Approximately one-half of the total primary production on Earth occurs in the oceans[127]. Prochlorococcus and Synechococcus are among the most abundant members of the cyanobacterial lineage and are important contributors to a significant fraction of the primary production and to global carbon cycling in the world's oceans[113].

For more than two decades now, the field of microbial ecology is being transformed by the emergence of new methodologies that enable assessments of the diversity and physiology of microbes in their natural habitat. Less than a decade ago, one of the main challenges was to link specific microbial groups to their activities *in situ*[23]. To do so, researchers have used FISH with microautoradiography[12] to track the incorporation of radiolabelled substrates into phylogenetically identified cell types. Phylogenetic identification has also been coupled with stable isotope tracers (for instance, 13C-enriched substrates) to identify substrates consumed by specific microbes[89]. Today, the establishment of environmental genomics, transcriptomics and proteomics will considerably extend our knowledge on microbial assemblages and their physiologies.

A grand challenge is to decipher the mechanisms regulating the operation and maintenance of elemental cycling on Earth by the microorganisms. To achieve this goal, environmental sequencing is an indispensable tool because the information encoded in microbial genomes significantly drives ecosystem processes. Biogeochemical and ecosystem processes are partly regulated by ecological interactions and community metabolism of microbial assemblages. One important goal is to identify and study the core set of genes responsible for fluxes of key elements in the context of a global metabolic pathway[35]. This is relevant in the context of climate change (e. g., ocean acidification): climate factors can select microbial populations and therefore influence the structure and maintenance of microbial communities. For example, it has been shown that environmental factors can predict annually recurring bacterial communities[43]. Therefore, it might be possible to monitor and survey the effects of climate change by measuring microbial diversity and composition, at least in the ocean. Also, it is crucial to understand the biological mechanisms that regulate carbon exchanges between the land and atmosphere, and how these exchanges respond to climate change[7].

Importantly, microbial interactions have to be considered for any project in microbial ecology, and recent research has shown that bacteria can communicate with one another using chemical signal molecules. The process of producing, releasing and detecting these molecules (for example in 'quorum-sensing'[125]) allows bacteria to monitor their environments and to synchronize their activities. At the community level (e. g., in biofilms), it is clear that quorum-sensing plays an important role by which microbes can communicate, cooperate and differentiate[82]. Intra- and interspecies signaling in microbes constitute a new important field of research that will help us reveal and understand the complex interactions between naturally co-occurring microorganisms.

Fortunately, the development of high-throughput sequencing and functional genomics is also revolutionizing the field of microbial ecology and can help biologists to characterize in great details the ecological relationships between coexisting microbes. A recent work, integrating proteomic, geochemical, and biological information from several microbial communities collected from an acid mine drainage environment, demonstrated the power of this strategy to investigate ecological and evolutionary relationships between microbial physiology and the environment[69].

In the following chapter, I will introduce the strategies that I have developed to study microbial life at the molecular and community levels in various environments and using a variety of technologies.

3 APPLIED ENVIRONMENTAL OMICS

3.1 MICROBIAL DIVERSITY CENSUS

The emergence of new sequencing techniques is changing the way we can measure the diversity of microbes *in situ*. With the development of the pyrosequencing technology and the 454 platform, it is now possible to analyze multiple samples in parallel. Mitchell Sogin and collaborators were the first to adapt the technology for 16S analysis[106]. They used PCR-amplification of a short hypervariable region of the 16S rRNA gene from distinct environments using universal primers, containing a unique sequence tag to identify each specific sample[81], and sequenced them simultaneously using a single 454 run. Using this approach, they generated thousands of short barcoded reads of more than 100 bp on average, which corresponds to a considerable increase in throughput compared to any Sanger-based method. By adopting this strategy, they could show that deep sea bacterial communities are one to two orders of magnitude more complex than previously reported for any microbial environment[106]. Hence, the deep sequencing of PCR amplicon libraries can enhance the detection of low-abundant populations in complex microbial ecosystems. Today, the third generation of pyrosequencing is already available and allows the production of roughly 500 Mb of reads per run with an average read length of more than 400 bp[114]. These enhancements in performance, throughput and cost will greatly improve the resolution and accessibility of this approach, which might ultimately replace the Sanger sequencing methods.

Importantly, it has to be noted that barcoding approaches are limited by the PCR-amplification step of the procedure which uses "universal" primers. Indeed, these primers are not truly universal and do not allow the recovering of certain microbial taxa[6]. Also, it is increasingly recognized that diversity can be over-estimated using this type of approach[58], but methods, protocols and softwares[58,88,16,102] are being developed to limit these problems and to facilitate the analyses. Certainly, given the deluge of information produced by these technologies, sequence data can no longer be analyzed manually, new tools and software need to be developed. On the data side, new efforts are on-going to standardize sequence meta-data (e. g., collection date, geographical position, temperature, etc.)[53] and this is really important in order to be able to connect microorganisms to their environment in space and time.

As of today, numerous studies have successfully employed barcoded, multiplexed pyrosequencing in various environments and revealed unprecedented diversity, evolutionary and ecological insights. Many interesting studies focused on the human gut microbiota which is composed of essential commensal bacteria[134,29,116]. We just begin to understand the role and the complex interactions between us and our "third genome" (i. e., we possess three distinct genomes: nuclear, mitochondrial and microbial). A striking example is the brain–gut–enteric microbiota axis: the brain can

indirectly influence commensal organisms (e. g., via signaling molecules released into the gut lumen) and communication from enteric microbiota to the host can also occur (e. g., through direct stimulation of host cells). Enteric microbiota are supposed to be directly or indirectly influenced, via changes in their environment or via host-enteric microbiota, respectively. This can be caused by the 'emotional motor system', which refers to several output systems (such as the hypothalamus–pituitary–adrenal axis or endogenous pathways that modulate pain and discomfort) that mediate the effect of emotional states on various bodily systems, including gastrointestinal function[93]. Microorganisms in the gut can also interact and influence mammalian cells via various signaling molecules: peptides, epidermal growth factor and autoinducers. Signaling molecules used for communication by vertebrates, invertebrates and microbes actually share structural similarities[96,65]. The mammalian noradrenergic signaling system shares homology with the microbial autoinducer 3–QseC signaling system. This causes the QseC system to be activated also by a specific host hormone and enables interkingdom signaling, relevant to brain–gut interactions during stress[93]. Indeed, bacteria use quorum-sensing to regulate gene expression in response to signals from other bacteria but also in response to host signals, and these mechanisms can mediate diverse physiological functions (e. g., secondary metabolite production, pathogenecity). It is suggested that these interactions might have an important role in modulating gut function and general well-being[93]. The human gut microbiota can directly influence our health, therefore it is important to understand its development[80], identity[116], stability[30] and its mammal-associated co-evolution[61].

In order to analyze the possible mechanisms of microbiota-mediated protection against pathogens (here, *Salmonella enterica*), we assessed the microbial diversity in the intestinal ecosystem of an infection model organism (i. e., a mouse, *Mus musculus*, with reduced gut microbial diversity) under various conditions, using barcoded parallel 16S rRNA gene sequencing. The results of this study are presented in Chapter 4. This work was initiated after a first collaboration, which also included the analysis of lineage abundance in various samples, and which resulted in another publication included in Appendix A.

3.2 SYSTEMS MICROBIOLOGY

Microorganisms and microbial communities constitute a good model for systems biology because they are easy to manipulate, non-expensive to cultivate and play crucial roles in the biosphere and human health. Various 'omics' technologies are now available to measure molecules at all levels of a cell's organization. Presently, DNA sequences can be determined (genomics), transcripts levels can be measured (transcriptomics), metabolites can be detected (metabolomics) and also proteins can be catalogued and quantified (proteomics). A great challenge is the integration of these various types of datasets (often very large) in order to understand how a microbial cell or community functions[52].

An interesting approach to characterize a given microbial ecosystem is the acquisition of information for both genomic sequences (what is encoded) and protein sequences (what is expressed). By this means, recent studies by Lo, Denef and co-workers have

gained insights into the ecological and evolutionary processes that shape microbial consortia in a low-complexity ecosystem. They developed a new approach to map identified peptides from various samples onto a nearly complete genome recovered from environmental genomics data[62]. The method allows to genotype the dominant bacteria (*Leptospirilum group II*) populations from several biofilm samples from an acid mine drainage cave. Using this approach, the authors identified six distinct genotypes that are recombinants, including segments from two 'parental' genotypes and indicating selection for distinct recombinants[25]. These findings indicate that recombination is an important mechanism for the fine-scale adaptation of this bacterial group within this eco-system. They also further described this process for ecological fine-tuning adaptation and demonstrated the main influence of environmental parameters (such as temperature) on the regulation of shared genes and the expression of a small subset of genes unique to each genotype, leading to distinct ecological strategies[26].

New analytical tools and methods need to be developed for the integration of data monitoring molecular activities of whole microbial communities. In this context, we collaborated with microbiologists interested in characterizing the eco-physiology of microbes inhabiting the surface of plant leaves (i. e., *the phyllosphere*). To do so, we developed a new method to estimate encoding and expression levels, from genomic and proteomic data, and to assign the detected functional capabilities to taxa present in the microbial community. This work is presented in Chapter 5 and discussed in Section 8.2.

Systems microbiology is aiming at completely characterizing microbial ecosystems. The importance of ecological interactions among microorganisms is increasingly recognized and they need to be included in a global framework that will help us to develop an understanding of community eco-system function[90].

3.3 COEXISTENCE AND EVOLUTION

To understand the ecology and evolution of microbes and communities, the first step is the description of associations between organisms and their environments. Thus, there is a need to globally monitor and understand inter-relationships among microbes[42]. Metagenomics technology is considerably broadening the study of microbial diversity, and new methods are now available to phylogenetically assign DNA fragments from the environment onto a tree of life[123,67]. The latter approach has already revealed that microbes have preferred habitats, where they are frequently detected, and that these preferences seem to change only slowly at evolutionary timescales[123].

Therefore, we can ask ourselves whether microorganisms also have preferred community partners, a question that has not been explored via the analysis of environmental sequence data, until recently. Known cases of partnerships involving microorganisms usually feature higher eukaryotes as hosts[131,132], but it is very likely that partnerships also occur specifically between single-celled organisms, as has been previously reported for some specific bacteria, albeit rather anecdotally[91,78]. Pathogens can also

form communities, particularly in chronic infections, affecting humans more severely. In this context, cataloguing microbial (co-)existence might lead to crucial insights and reveal new microbial ecosystems or complex metabolizing capabilities that may even be biotechnologically relevant.

To tackle this question, we developed an *in silico* pipeline to globally identify specific associations among bacteria and archaea using publicly available environmental sequence data. The methods and results of our analysis are presented in Chapter 6.

Part II

RESULTS

4 ASSESSING MICROBIAL DIVERSITY USING HIGH-THROUGHPUT SEQUENCING

4.1 PREFACE

The commensal microbiota inhabiting the intestinal tract is fundamental to human health. Notably, it contributes to protection against gastrointestinal infections by pathogens. So far, the mechanisms responsible for this colonization resistance are largely unknown. Here, using a mouse model for *Salmonella enterica* induced gut inflammation and microbiota analysis by 454 barcoded-amplicon pyrosequencing, we revealed characteristics of the commensal microbiota indicative for a high or low degree of colonization resistance. Our analyses led to the description of a new concept in gut ecosystem biology.

For this study[112], I performed all the sequence data analyses (sequence clustering, abundance profile clustering and quantitative co-occurrence analyses) in order to create the figures 3, 4 and 6 of the publication.

4.2 LIKE WILL TO LIKE

The publication is included below.

OPEN ACCESS Freely available online

PLoS PATHOGENS

Like Will to Like: Abundances of Closely Related Species Can Predict Susceptibility to Intestinal Colonization by Pathogenic and Commensal Bacteria

Bärbel Stecher[1☯]*, Samuel Chaffron[2☯], Rina Käppeli[1☯], Siegfried Hapfelmeier[3,4], Susanne Freedrich[1], Thomas C. Weber[1], Jorum Kirundi[3,4], Mrutyunjay Suar[5], Kathy D. McCoy[4], Christian von Mering[2], Andrew J. Macpherson[3,4], Wolf-Dietrich Hardt[1]

1 Institute of Microbiology, ETH Zürich, Zürich, Switzerland, **2** Institute of Molecular Biology and Swiss Institute of Bioinformatics, University of Zürich, Zürich, Switzerland, **3** Gastroenterology Inselspital, Department Klinische Forschung, Bern, Switzerland, **4** Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada, **5** School of Biotechnology, KIIT University, Bhubaneswar, Orissa, India

Abstract

The intestinal ecosystem is formed by a complex, yet highly characteristic microbial community. The parameters defining whether this community permits invasion of a new bacterial species are unclear. In particular, inhibition of enteropathogen infection by the gut microbiota (= colonization resistance) is poorly understood. To analyze the mechanisms of microbiota-mediated protection from *Salmonella enterica* induced enterocolitis, we used a mouse infection model and large scale high-throughput pyrosequencing. In contrast to conventional mice (CON), mice with a gut microbiota of low complexity (LCM) were highly susceptible to *S. enterica* induced colonization and enterocolitis. Colonization resistance was partially restored in LCM-animals by co-housing with conventional mice for 21 days (LCM^{con21}). 16S rRNA sequence analysis comparing LCM, LCM^{con21} and CON gut microbiota revealed that gut microbiota complexity increased upon conventionalization and correlated with increased resistance to *S. enterica* infection. Comparative microbiota analysis of mice with varying degrees of colonization resistance allowed us to identify intestinal ecosystem characteristics associated with susceptibility to *S. enterica* infection. Moreover, this system enabled us to gain further insights into the general principles of gut ecosystem invasion by non-pathogenic, commensal bacteria. Mice harboring high commensal *E. coli* densities were more susceptible to *S. enterica* induced gut inflammation. Similarly, mice with high titers of Lactobacilli were more efficiently colonized by a commensal *Lactobacillus reuteri* ^{RR} strain after oral inoculation. Upon examination of 16S rRNA sequence data from 9 CON mice we found that closely related phylotypes generally display significantly correlated abundances (co-occurrence), more so than distantly related phylotypes. Thus, in essence, the presence of closely related species can increase the chance of invasion of newly incoming species into the gut ecosystem. We provide evidence that this principle might be of general validity for invasion of bacteria in preformed gut ecosystems. This might be of relevance for human enteropathogen infections as well as therapeutic use of probiotic commensal bacteria.

Citation: Stecher B, Chaffron S, Käppeli R, Hapfelmeier S, Freedrich S, et al. (2010) Like Will to Like: Abundances of Closely Related Species Can Predict Susceptibility to Intestinal Colonization by Pathogenic and Commensal Bacteria. PLoS Pathog 6(1): e1000711. doi:10.1371/journal.ppat.1000711

Editor: Howard Ochman, University of Arizona, United States of America

Received September 18, 2009; **Accepted** November 25, 2009; **Published** January 8, 2010

Funding: This work was supported by grants to WDH from the Swiss National Science Foundation (310000-113623/1), the European Union (SavinMucoPath No. 032296) and ETH Zürich Research foundation (TH-08 08-3). Sequencing was supported by Genome Canada and the Canada Research Council. BS was supported by the UBS AG (Zurich) on behalf of a customer and the European Molecular Biology Organization (Embo short-term Fellowship). Work was also supported by the University of Zürich Research Priority Program in Systems Biology/Functional Genomics. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: stecher@micro.biol.ethz.ch

☯ These authors contributed equally to this work.

Introduction

The mammalian intestine hosts a microbial community of astonishing density and complexity. This intricate association presumably required significant coevolution of the host and its microbiota. Apparently, this coevolution has been guided by positive selection for factors that result in a state of both mutual tolerance and benefit.

Microbial colonization of the intestine takes place right after birth and complexity steadily increases henceforward. The temporal and spatial assembly of the gut microbiota is apparently not guided by specific rules but eventually, after weaning, a stable microbial ecosystem is formed [1]. The adult human intestine hosts 10^{13} to 10^{14} bacteria belonging to at least 500 different species or strains [2]. Up to 9 different bacterial phyla are usually found; however, the Firmicutes and Bacteroidetes account for over 90% of all bacteria [3]. Despite its striking conservation on a higher phylogenetic level, the abundance of bacteria on species or strain level varies extensively between non-related individuals. Nevertheless, a core gut microbiome (= sum of microbial genes) that is shared among different individuals ensures conservation of metabolic functions provided by the microbiota [4]. It is assumed that the microbial ecosystem, once it is formed, efficiently prevents invasion by foreign species. This has been extensively studied in

Author Summary

The commensal microbiota, populating the intestinal tract to high levels, is fundamental to human health. It exerts beneficial effects on the immune system and contributes to protection against gastrointestinal infections (= colonization resistance) by largely unknown mechanisms. Here, we reveal characteristics of the commensal microbiota indicative for a high or low degree of colonization resistance. Using a mouse model for *Salmonella enterica* induced gut inflammation and microbiota analysis by 454 amplicon sequencing, we show that mice having different types of microbiota exhibit differential susceptibility to pathogen infection. In addition, our data lead to the description of a new concept in gut ecosystem biology: the intrusion-success of an extrinsic bacterial species into an established gut ecosystem is related to the abundance of closely related bacteria, already present in this gut ecosystem. We show that this principle applies not only to enteropathogen infection but also to inoculation with beneficial gut bacteria. Humans can display largely different degrees of susceptibility to enteric infections. Similarly, the effectiveness of probiotic therapy varies greatly from person to person. Our data might explain these differences and could be used for increasing the efficacy of probiotic therapy and for identifying patients at risk of developing enteric infections.

the case of enteric pathogens and is known as 'colonization resistance' (CR) [5].

The gut microbiota protects its host against infection by life-threatening pathogens such as *Vibrio cholerae*, pathogenic *Escherichia coli* strains, *Shigella* spp., *Clostridium difficile* and *Salmonella* spp. [6,7]. To date, the molecular bases of CR as well as the key bacterial species involved remain poorly defined. It is clear that if the gut microbiota is absent or disturbed (i.e. germfree status, antibiotic treatment, gut inflammation) the infection risk increases drastically [8,9,10,11,12]. CR might not only exclude pathogenic bacteria but also acts against harmless or even beneficial bacteria, such as probiotics. For example, the efficiency of probiotic therapy can differ greatly among individuals [13,14,15,16]. To increase effectiveness of probiotic therapy, research aims at improving the half-life of probiotic strains in the gut [17].

In this study we set out to identify characteristics of the bacterial gut microbiota that are linked to infectivity of the human pathogen *Salmonella enterica*. Conventional mice (CON) harbouring a complex gut microbiota are highly resistant to oral *Salmonella enterica* infection and concomitant induction of gut inflammation [18]. We tested colonization resistance of mice harbouring different types of gut microbiota. On a quantitative level, we found that mice having a higher gut microbiota complexity exhibited increased protection against *Salmonella*-induced gut inflammation. In addition we found that the invasion-success of novel species into an established gut ecosystem (i.e. *Salmonella enterica*, *Lactobacillus reuteri* ^{RR}) may be predetermined by the abundance of species that are closely related to the invader.

Materials and Methods

Animals

We generated LCM mice by colonizing germfree mice with the Altered Schaedler flora (ASF) according to the protocol published on the Taconic webpage. Mice were inoculated at eight weeks of age by intra-gastric and intra-rectal administration of 10^7–10^8 c.f.u. of ASF bacteria on consecutive days (www.taconic.com/library). LCM mice (C57Bl/6 background) were maintained under barrier conditions in individually ventilated cages with autoclaved chow and autoclaved, acidified water. No mice with complex gut microbiota were housed in the same room to prevent contamination with natural gut bacteria. CON C57Bl/6 mice were obtained from Janvier (France), Charles River Laboratories (Sulzfeld, Germany), from the Rodent Center HCI (RCHCI Zürich) and the Biologisches Zentrallabor (BZL; Univeristy Hospital Zurich). CON transgene negative B6.129P-*CX3CR1*^{tm1Litt}/J mice (CX3CR1) [19] and CON Ly5.1 (B6.SJL-*Ptprc*^{a} *Pepc*^{b}) were bred at the RCHCI Zürich and CON heterozygous MyD88$^{+/-}$ mice (C57BL/6 background) [20] at RCC Füllinsdorf, respectively. All mice were bred and kept specified pathogen free in individually ventilated cages. This restricts microbial transfers between mice housed in the same room and animal facility.

LCM mice, CON mice or streptomycin-pretreated CON mice (20 mg/animal 24h prior to *Salmonella* infection) were infected by gavage with 5×10^7 CFU *S.* Typhimurium SL1344 wildtype or avirulent (*sseD*::*aphT* [21]) strains or *S.* Enteritidis 125109 (streptomycin-resistant variant M1525 [22]). Live bacterial loads in mesenteric lymph nodes (MLN), spleen and cecal content were determined by plating on MacConkey-agar (Oxoid) with respective antibiotics [21]. *Lactobacillus reuteri* ^{RR} ($8*10^6$ cfu i.g.) was administered by gavage and cultured anaerobically on MRS media (Biolife; 100 µg/ml rifampicin). To enoumerate bacteria, cecal content was stained with Sytox-green and bacteria were counted in a Neubauer-chamber. Bacterial density is given as Sytox-green positive bacteria per gram cecal content.

Ethics statement

All animal experiments were approved (license 201/2004 and 201/2007 Kantonales Veterinäramt Zürich) and performed as legally required.

Bacteria

The streptomycin-resistant wild type strain *S.* Typhimurium (SL1344 wildtype [23]), the isogenic mutant *S.* Typhimurium ^{avir} (Δ*invG sseD*::*aphT*; *kan*^{R} [24]) and wild type *S.* Enteritidis (M1525 [22]) were grown in LB 0.3 M NaCl as described [24]. *L. reuteri* ^{RR} [12] was isolated from our mouse colony selected on MRS media (100 µg/ml rifampicin) (Biolife) and grown anaerobically.

Histology

HE-stained cecum cryosections were scored as described, evaluating submucosal edema, PMN infiltration, goblet cells and epithelial damage yielding a total severity score of 0-13 points [21]. 0–3 = no to minimal signs of inflammation which are not sign of a disease; this is frequently found in the cecum of conventional mice. 4–8 = moderate inflammation; 9–13 = profound inflammation.

Statistical analysis

Statistical analysis of *Salmonella* colonization titers was performed using the exact Mann-Whitney U Test (SPSS Version 14.0). P-values less than 0.05 (2-tailed) were considered statistically significant. Pearson- and Spearman correlation coefficients for bacterial colonization levels were calculated using Graphpad Prism (Version 5.01). Other statistical analyses (Pearson correlation, Kolmogorov-Smirnov test) were performed using the statistical language and environment R (http://www.r-project.org/). To systematically detect differentially abundant OTUs in all mice and for different clustering distances, we used the R software Metastats [25].

Bacterial DNA extraction and 16S rRNA gene specific PCR

Total DNA was extracted from cecal contents using a QIAmp DNA stool mini kit (Qiagen). Bacterial lysis was enhanced using 0.1 mm glass beads in buffer ASF and a Tissuelyzer device (5 minutes, 30 Hz; Qiagen). V5-V6 regions of bacterial 16S rRNA were amplified using primers B-V5 (5′ GCCTTGCCAGCCC-GCTCAG ATT AGA TAC CCY GGT AGT CC 3′) and A-V6-TAGC (5′ GCCTCCCTCGCGCCATCAG [TAGC] ACGA-GCTGACGACARCCATG 3′). The brackets contain one of the 20 different 4-mer tag identifiers [TAGC, TCGA, TCGC, TAGA, TGCA, ATCG, AGCT, AGCG, ATCT, ACGT, GATC, GCTA, GCTC, GATA, GTCA, CAGT, CTGA, CAGA, CTGT, CGTA;]. Cycling condition were as follows: 95°C, 10 min; 22 cycles of (94°C, 30 s; 57°C, 30 s; 72°C, 30 s); 72°C, 8 min; 4°C, ∞; Reaction conditions (50 µl) were as follows: 50 ng template DNA; 50 mM KCl, 10 mM Tris-HCl pH 8.3, 1,5 mM Mg^{2+}, 0,2 mM dNTPs; 40 pmol of each primer, 5U of Taq DNA polymerase (Mastertaq; Eppendorf).

PCR products of different reactions were pooled, ethanol-precipitated and fragments of ~300 bp were purified by gel electrophoresis, excised and recovered using a gel-extraction kit (Machery-Nagel). Amplicon sequencing of the PCR products was performed using a 454 FLX instrument (70×70 Picotitre plate) according to the protocol recommended by the supplier (www.454.com). PCR to detect ASF bacteria in the feces was done as described in [26].

E. coli differentiation

Candidate *E. coli* strains yielding large, red colonies on MacConkey agar were typed using Enterotubes (BD Biosciences). Additionally, in some cases 16S rRNA gene sequencing was performed. The amplification was performed with extracted DNA using "broad-range" bacterial primers fD1 and rP1 [27]. Reaction conditions were as follows: Deoxyribonucleoside triphosphates (0.25 mM), primers (1 pmol/µl each), 5UTaq-DNA polymerase (Mastertaq; Eppendorf), 50 ng of template DNA. The following cycling parameters were used: 5 min of initial denaturation at 94°C followed by 35 cycles of denaturation (1 min at 94°C), annealing (1 min at 43°C), and elongation (2 min at 72°C), with a final extension at 72°C for 7 min. Amplified PCR products were purified by gel electrophoresis and sequenced using rP1 as sequencing primer. Sequences were assigned to the RDP taxonomy using the RDP classifier (http://rdp.cme.msu.edu/; [28]).

Quantification of Lactobacilli

Fecal samples were re-suspended in PBS and plated in appropriate dilutions on MRS agar (DE MAN, ROGOSA und SHARPE; Biolife) that supports growth of *Lactobacillus* spp. as well as *Leuconostoc* spp. and *Pediococcus* spp. Plates were incubated for 24 h in an atmosphere of 7% H_2, 10% CO_2 and 83% N_2 at 37°C in anaerobic jars.

Reads sorting and quality filtering

The amplicon library was sequenced according to the 454 Amplicon Sequencing protocols provided by the manufacturer (Roche 454) at the McMaster University Hamilton (Canada). The sequence determination was made using GS Run Processor in Roche 454 Genome Sequencer FLX Software Package 2.0.00.22. Performance of the sequencing run was gauged using known pieces of DNA introduced in the sequencing run as DNAControl Beads. On average, 94% of reads from DNA Control Beads matched the corresponding known sequences with at least 98% accuracy over the first 200 bases, which was above the typical threshold (80% matches of 98% accuracy over 200 bases). To estimate the reliability of sample separation using our primer-tagging approach, we assessed the number of reads observed to have an illegitimate 4-mer tag (i.e., different from our set of 20 tags). The sequencing plate (including other non-analyzed samples) produced a total of 264,503 reads from which 1,339 contained a wrong tag (0.506%). Given that 256 distinct 4-mer tags are possible and that we used only 20 of these, the majority of sequencing errors in this region are detectable. Correcting for the small fraction of undetectable errors (20/256) and division by four yields a sequencing error rate of 0.137% per single nucleotide - at the position of the tag in the primer (this includes errors during primer synthesis as well as sequencing). Because most errors are actually visible as errors, the rate of unintentional 'miscall' of the sample is 0.043%.

We applied quality control of 454 reads in order to avoid artificial inflation of ecosystem diversity estimates [29]. Reads containing one of the exact 4 nt tag sequences were filtered with respect to their length (200 nt ≤ length ≤300 nt). Quality filtering was then applied to include only sequences containing the consensus sequence ('ACGAGCTGACGACA[AG]CCATG') of the V6 reverse primer and displaying at maximum one ambiguous nt 'N'. The latter criterion has been reported as a good indicator of sequence quality for a single read [30]. We identified 5,268 reads shorter than 200 nt, 228 reads longer than 300 nt and 2,169 reads containing more than one 'N'. After filtering, 190,728 reads remained (initial total of 197,949 reads containing the exact primer sequence and tag) and were processed as described below.

Definition of OTUs

OTUs were defined using the complete filtered dataset, with the exception of exactly identical reads, which were made non-redundant to reduce computational complexity. Before OTU generation, we added reference sequences for subsequent taxonomic classification of OTUs; for this, we used a reference database of selected 16S rRNA gene sequences downloaded from the Greengenes database (http://greengenes.lbl.gov/Download/Sequence_Data/Greengenes_format/greengenes16SrRNAgenes.txt.gz, release 01-28-2009 [31]). In Greengenes, all entries are pre-annotated using several independent taxonomy inferences including the RDP taxonomy. Our reference database was built using full-length non-chimeric sequences with a minimum length of 1100 nt (in order to fully cover the V6 region of all entries). No archaeal sequences were included.

The alignment of non-redundant reads from all mice with the reference database was performed using the secondary-structure aware Infernal aligner (http://infernal.janelia.org/, release 1.0, [32]) and based on the 16S rRNA bacterial covariance model of the RDP database (http://rdp.cme.msu.edu/; [28]).

Before defining OTUs, we first removed reference sequences for which the alignment was not successful (Infernal bitscore <0). The alignment was then processed to include an equivalent amount of information from every read. To do so, we identified the consensus reverse primer sequence of the V6 region within the aligned sequence of *Escherichia coli* K12, as a reference. The full alignment was then trimmed from the start position (defined by the *E. coli* V6 reverse primer) and ended after 200 nt's. This also insured the limitation of the effect of pyrosequencing errors by trimming the 3′ end of each read, a region which is more sequencing-error prone (the trimmed and aligned reads length ranged from 192 to 241 nt) [29]. Using this alignment, OTUs were built by hierarchical cluster analysis at various distances (0.01, 0.03, 0.05, 0.10, 0.15

and 0.2) using the 'complete linkage clustering' tool of the RDP pyrosequencing pipeline (http://pyro.cme.msu.edu/ [28]).

Taxonomy assignment

As a first step, taxonomy was predicted for all reads using the stand-alone version of the RDP classifier (http://sourceforge.net/projects/rdp-classifier, revision 2.0, [33]). Taxon predictions were considered reliable if supported by a minimum bootstrap value of 80%. In order to predict taxonomy for each OTU, we either used any reference sequences present within a cluster, or the taxonomy of the reads present in the cluster, as predicted by the RDP classifier. To increase the resolution of the prediction, we privileged any reference sequences over the reads. For each OTU, taxonomy was inferred by a simple majority vote: if more than half of the reference sequences (or reads) present within a cluster agreed on a taxon, the OTU was annotated according to this taxon. In case of conflicts, we assigned a consensus taxon to a higher phylogenetic level for which the majority vote condition was respected.

OTU distribution between the different experimental groups and predicted taxonomies were visualized as heatmaps generated by custom Python scripting and the statistical software package R (www.r-project.org).

Chimera estimation

Deep pyrosequencing on the 454 platform has revealed extensive microbial diversity that was previously undetected with culture-dependent methods [34]. Nevertheless, the details of protocols to generate this type of data should always be carefully considered; various types of bias can be introduced at different steps. Here, sequencing was performed on pools of PCR products, thus limitations and biases of this technique have to be taken into account to interpret the results. The abundances of amplicons may not accurately reflect the relative abundances of the template DNA because of differential primer binding- and elongation-efficiencies. Moreover, during amplification, chimeric sequences can be generated.

On such short sequences, recombination points (recombination can occur from an incompletely extended primer or by template-switching [35]) are extremely difficult to detect. Recently, a new tool to filter noise and remove chimera in 454 pyrosequencing data has been published [36]. In this study, the authors suggest that because of sequencing errors, diversity estimates may be at least an order of magnitude too high. To our best knowledge, at the time of analysis, there were no available tools to detect chimera within libraries of short 454 reads. Therefore, to detect chimera we decided to compare taxonomies assigned to N-terminal and C-terminal read fragments. A read was regarded as 'non-chimeric' if the best hits (BLASTn) for both of its fragments had a minimum identity of 95% and a minimum bit-score of 150. These cutoffs were selected heuristically in order to insure a reasonable alignment length and a relatively high identity to the matching reference sequence. A given read was deemed chimeric when the taxonomies of the best hits of each half were clearly not congruent (i.e., differing at the phylum level). Our simple chimeric reads detection method resulted in a higher rate of detected chimera compared to the method of Quince et *al.*, 2009 (~7% compare to ~3% in their example) adapted from the Mallard algorithm [37], suggesting that our approach is probably stringent enough at detecting chimera [36].

OTU abundance correlation analysis

In order to test the general hypothesis that closely related bacteria are present at similar levels in CON mice, we systematically compared the relative abundance between all OTUs detected in 9 distinct CON mice. Here, a detected OTU was defined as present in at least 6 mice (2/3). For each possible pair of OTUs, we computed the Pearson correlation coefficient of their relative abundance (number of reads normalized by the total number of reads in a given sample) in each CON mouse. To compare these results to the distance between 2 OTUs we computed identities between all considered OTUs using their representative sequences in the complete alignment (all reads and all reference sequences). An OTU's representative sequence is defined as the sequence that has the minimum sum of the square of the distances to all other sequences within that cluster.

For statistics inference, we semi-randomized our results by shuffling non-null abundances between all detected OTUs. For both distributions we plotted running medians (y-axis) with a window size of 500 data points (the window size was decreased towards the beginning and end of the distributions). The Kolmogorov-Smirnov test (one- and two-tailed) was used to compare both distributions (actual data and random data) with respect to the deviation of the running median from the random expectation. The test was computed on x-axis bins (0.1) in order to better interpret the results of the analysis. The data processing and plotting were performed using Python scripting.

Results

LCM mice are susceptible to *S.* Typhimurium induced gut inflammation without the need for antibiotic treatment

Germfree mice and CON mice orally treated with a single dose of antibiotic (i.e. aminoglycosides, β-lactams, vancomycin) are highly susceptible to enteric *S.* Typhimurium colonization and develop acute inflammation of the lower intestine (cecum, colon) upon oral infection [11,18,38]. Here, we tested susceptibility of gnotobiotic mice, associated with a standardized low complex type of gut microbiota (termed LCM), to oral *S.* Typhimurium infection. In contrast to CON mice (~500 different bacterial strains in the gut), the gut microbiota of LCM mice includes a mixture of only 8 bacterial strains, the Altered Schaedler Flora (ASF), which are typically found in the gut of rodents [39]. In order to test whether LCM mice were susceptible to oral *S.* Typhimurium infection, we infected unmanipulated LCM mice (n = 5) by orally gavaging them with *S.* Typhimurium wild type (5×10^7 cfu). As control, we infected age-matched groups of CON mice (n = 5) harboring a normal fully differentiated gut microbiota and CON mice pretreated with streptomycin 24 h prior to infection (smCON). All mice were sacrificed at 3 days p.i. *S.* Typhimurium titers at in the mLN and spleen were highest in smCON mice, while no difference was observed comparing CON and LCM groups (Fig. 1A,B). In keeping with previous work, the cecum of untreated CON mice was poorly colonized by *S.* Typhimurium (below 10^5 cfu/g) while smCON mice displayed high *S.* Typhimurium levels in their gut ($>10^8$ cfu/gram; $p<0.05$; Fig. 1C). Interestingly, LCM mice also displayed high pathogen titers in the cecum. Owing to this high-level colonization, wild type *S.* Typhimurium triggered a fulminant inflammatory response in the cecum and colon of both smCON and LCM mice, while no pathological changes could be observed in the CON mice not pretreated with antibiotics (Fig. 1D,E; Fig. S1). This demonstrates that, in contrast to normal complex type of gut microbiota, colonization of mice with a LCM gut microbiota does not confer CR against *S.* Typhimurium.

To verify that mucosal inflammation induced by *S.* Typhimurium in infected LCM mice is induced by *Salmonella*-specific virulence factors, we infected LCM mice with an avirulent mutant

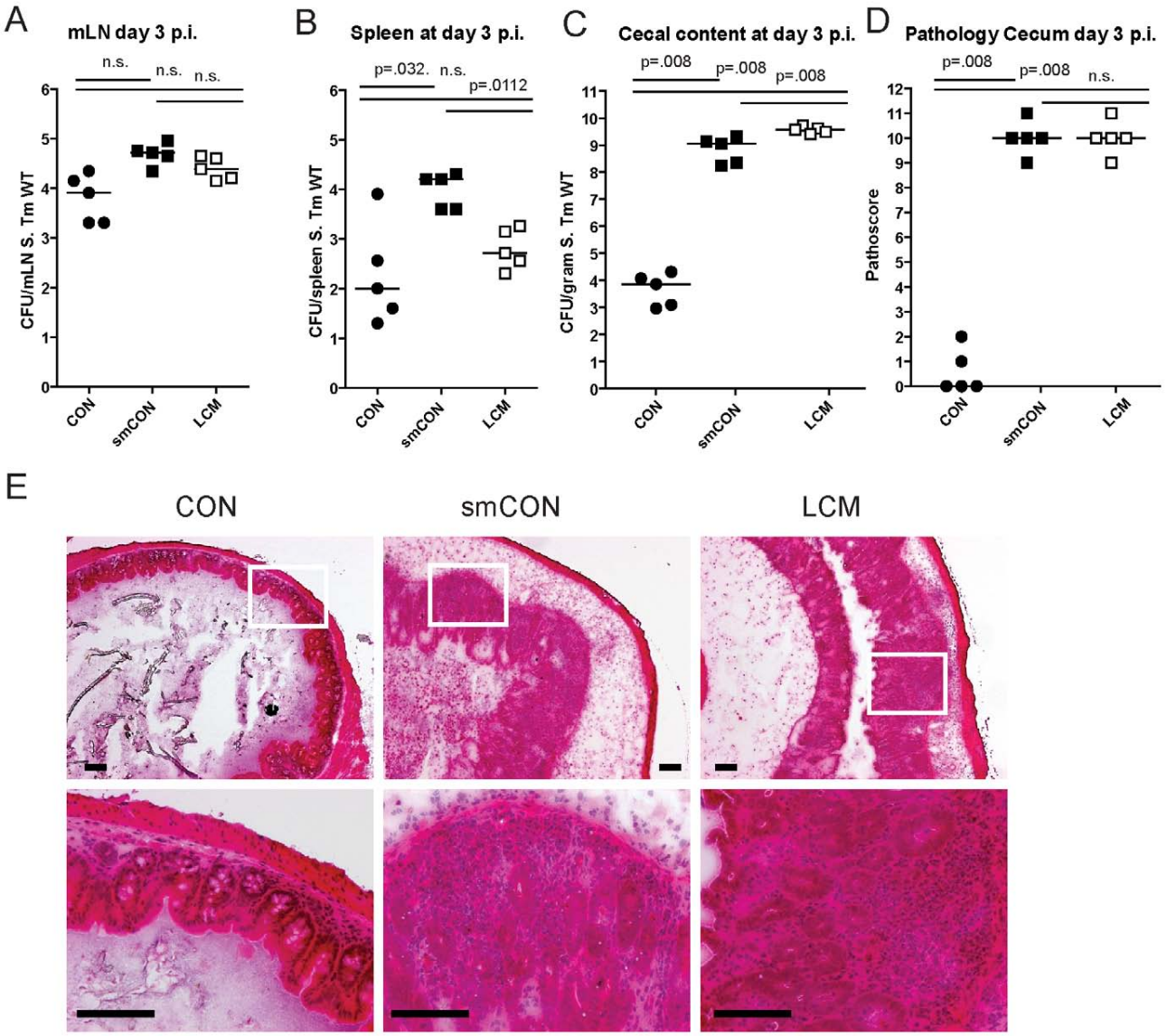


Figure 1. LCM mice susceptible to *S.* Typhimurium induced colitis. Groups (n = 5) of CON, streptomycin-treated mice (20 mg 24 h before infection) and LCM mice were infected with 5×10^7 cfu *S.* Typhimurium wild type by gavage and sacrificed at day 3 postinfection. S. Typhimurium levels in the mLN (A), spleen (B) and cecal content (C). (D) Cecal pathology scored in HE-stained tissue sections (see M&M). (E) HE-stained sections of cecal tissue from indicated mice. Enlarged section (white box) is shown in the lower panel. Scale bar: 100 µm.
doi:10.1371/journal.ppat.1000711.g001

lacking a functional TTSS-1 and 2 (*S.* Typhimurium^{avir}; 5×10^7 cfu). Despite colonizing the gut to high titers, *S.* Typhimurium^{avir} did not cause observable signs of intestinal pathology in LCM mice, demonstrating that gut inflammation in LCM mice was triggered by the same pathogenetic mechanisms as shown for smCON mice (Fig. S2).

Colonization resistance is transferred by re-association with a conventional gut microbiota

LCM mice, with a low complexity gut microbiota are susceptible to oral *S.* Typhimurium infection and develop severe acute colitis comparable to germfree or antibiotic-treated mice. Of note, microbiota in the cecum of LCM mice had a similar density as in CON mice (Fig. S3). These findings suggested that their gut microbiota lacks key bacterial species responsible for mediating CR. We reasoned that these protective bacteria would be transferable by co-housing LCM together with CON mice in the same cage. To test this hypothesis, we re-associated 2 groups of LCM mice (n = 2, 4) with one CON donor mouse each for 21 days. As controls, we used groups of non re-associated LCM and CON mice. We infected all animals with *S.* Typhimurium wild type (5×10^7 cfu by oral gavage) to measure the degree of CR.

Compared to unmanipulated LCM, all re-associated LCM mice had significantly lower *S.* Typhimurium loads in their feces at 1 day p.i. (Fig. 2A). 4 out of 6 animals were completely protected from *Salmonella*-colitis and did not show any signs of cecal pathology (Fig. 2E,F) while 2 out of 6 animals developed signs of inflammation (pathoscore 6 and 7) at day 3 p.i., which correlated with higher *S.* Typhimurium loads in the cecum of these mice (Fig. 2B). Systemic *S.* Typhimurium colonization appeared also

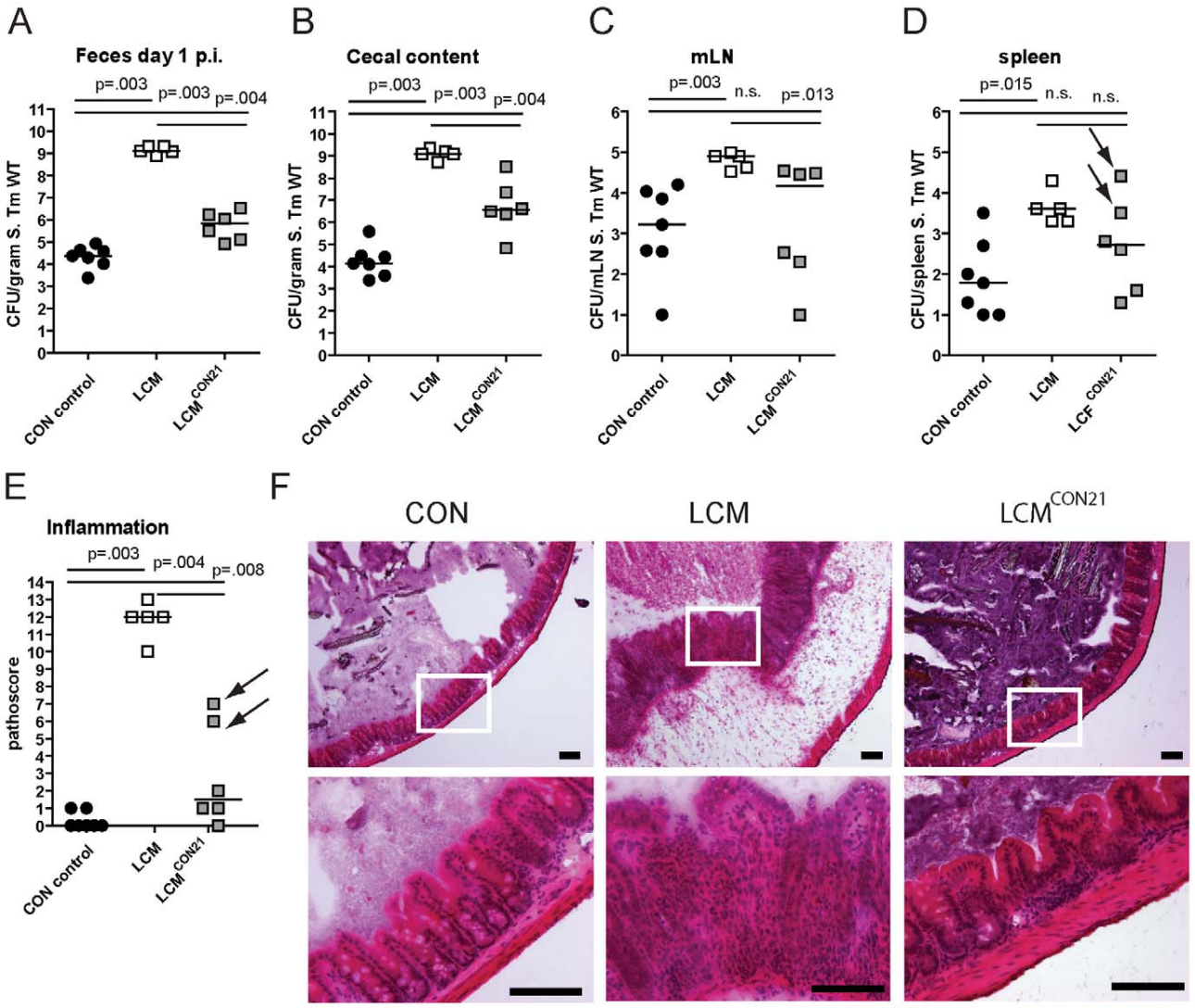


Figure 2. LCM gain CR by re-association with normal CON microbiota. Groups (n = 2,4) of LCM mice were re-associated with 1 CON donor each for 21 days in the same cage. Afterwards, non-reassociated LCM (control; n = 5), CON (control; n = 5) and re-associated LCM (n = 6) were infected with 5×10^7 cfu *S.* Typhimurium wild type by gavage for 3 days. *S.* Typhimurium levels in the feces at day 1 post infection (A), cecal content (B), mLN (C), spleen (D). (E) Cecal pathology scored in HE-stained tissue sections (see M&M). (F) HE-stained sections of cecal tissue from indicated mice. Enlarged section (white box) is shown in the lower panel. Scale bar: 100 μm. Arrows point at 2 mice that developed inflammation.
doi:10.1371/journal.ppat.1000711.g002

slightly reduced in re-associated LCM mice (Fig. 2C,D). This revealed that CR is transferable and suggested that discrete bacterial species transferred during the 3 week re-association contributed to colonization resistance and protection from colitis.

Microbiota analysis by high throughput amplicon-pyrosequencing

This offered the opportunity to correlate the changes in microbiota composition in the LCM mice with acquisition of colonization resistance. Protection from *Salmonella* diarrhea is conferred by bacteria entering the gut microbiota of LCM mice. To identify bacteria transferred during re-association, we analyzed gut microbiota composition by high-throughput sequencing of bacterial 16S rRNA genes. We analyzed the fecal microbiota because this non-invasive sampling method allows monitoring the microbiota of a given animal at various time points (i.e. before/after re-association or *Salmonella* infection). In contrast to other studies [2,4,34], we decided to sequence the 16S rRNA hypervariable regions V5 and V6 (length in *E. coli*: ~280 bp). Several studies have shown that sequencing of different hypervariable regions or full-length 16S rRNA genes yields to comparable results [30]. Thus we reasoned that V5V6 sequencing would not lead to a major bias in microbiota composition and at the same time would allow us to fully use current pyrosequencing capacity (the average output length of the 454FLX instrument is 250 bp).

After read-quality filtering, we obtained 190,728 reads with a length between 200–300 bps in total. Among those, 50,860 were non-redundant. The frequency of chimera, using a simple identification approach was 6.9% of the total reads (13,206) and 14.7% of non-redundant reads (7,499) (Fig. S4). This chimera-frequency is relatively high considering that we probably detected

only a fraction of chimeric reads using our method (materials and methods). Sequence reads were aligned with all quality-filtered sequences of our reference database generated from the Greengenes database [31] and operational taxonomic units (OTUs) were defined by hierarchical clustering at various distances, from 0.01 to 0.2. Taxonomy assignment was inferred using annotation from the reference sequences, if possible, or by predictions generated by the RDP classifier from the RDP database [28].

Microbiota complexity differs between LCM, LCM^{CON21} and CON mice

Comparing the average number of OTUs at various distances, clearly the CON donor mice display the highest level of complexity (Tables S1, S2 and S3). We found an average of 767±233 OTUs at a Clustering Distance (CD) of 0.03 and 499±139 OTUs at a CD of 0.05 (before chimera removal: 971±290 OTUs at a distance of 0.03 and 662±186 OTUs at a CD of 0.05). Complexity of the LCM gut microbiota was, as expected, relatively low. By strain-specific PCR [26], we only detected 4 members of the ASF (ASF361, ASF457, ASF500 and ASF519; Fig. S5). However, 29±10 OTUs at a 0.03 CD, and 17±5 OTUs at a 0.05 CD were detected (before chimera removal: 38±10 OTUs at a CD of 0.03 and 23±5 OTUs at a CD of 0.05). This was expected considering the way the LCM mice were generated. LCM status was created by inoculating germfree mice with bacteria of the ASF. Afterwards, LCM mice were kept in individually ventilated cages (IVCs). During this phase, a limited number of additional species might have been acquired. This might explain why our sequence analysis detected more than 8 different phylotypes in unmanipulated LCM mice. Alternatively, the relatively high number of phylotypes could be explained by PCR artifacts or most likely by the intrinsic error rate of pyrosequencing that can lead to a severe over-estimation of microbial diversity using the 16S rRNA marker gene [29]. In LCM^{CON21} mice, we observed a significant increase in gut microbiota complexity compared to LCM mice. At a 0.03 CD, 295±34 OTUs and at a CD of 0.05, 188±23 OTUs were detected (before chimera removal: 409±60 OTUs at a CD of 0.03 and 279±45 OTUs at a CD of 0.05). However, complexity in LCM^{CON21} mice remains significantly lower than that in CON mice.

We assessed the richness (actual diversity) of our samples by calculating the Shannon index (H) and species evenness (E) as well as the Chao1 diversity estimate (Tables S1, S2 and S3). These calculations revealed that the community was clearly under-sampled; for small CDs (0.01 to 0.05), the Chao1 estimator was, for each mouse, higher than the total number of OTUs. Although under-sampling is limiting our view on the true microbial diversity, it is legitimate to use diversity measures for relative comparisons among samples. Within this context, it is interesting to ask whether, after re-association, LCM mice display similar or different species evenness E compared to the CON mice. Here, species evenness can be regarded as the equilibrium between community members; the less variation is observed between species, the higher is the E value (in other words, evenness is greatest when species are equally abundant). The E-value is defined as the ratio of the theoretically maximal Shannon-index (if all observed phylotypes were equally abundant) divided by the actual Shannon-index. For a 0.05 CD, CON mice displayed an average E-value of 0.76 compared to an average of 0.70 for the LCM^{CON21} mice (compared to E = 0.15 for LCM mice). Thus, there is no major difference between CON mice and re-associated LCM mice with respect to evenness. Hence the 21 days of co-housing were sufficient in order to adopt a relatively complex and 'in equilibrium' microbial gut community.

To compare species richness between the 3 different groups, rarefaction curves were created for different CDs (Fig. 3; Fig. S6). For a CD of 0.01, slopes for CON and re-associated CON mice are rather steep, revealing again a considerable under-sampling in our experiment. However, slopes for 0.05 (for re-associated LCM) and 0.1 CD (for CON) seem to reach saturation, suggesting that for this level of analysis, the sampling was sufficiently complete. Therefore we decided to perform OTU analyses using a CD higher or equal to 0.05. This CD is in accordance with a recent report advising a stringent quality-based filtering of 16S- 454 reads and the use of a clustering threshold no greater than 97% [29].

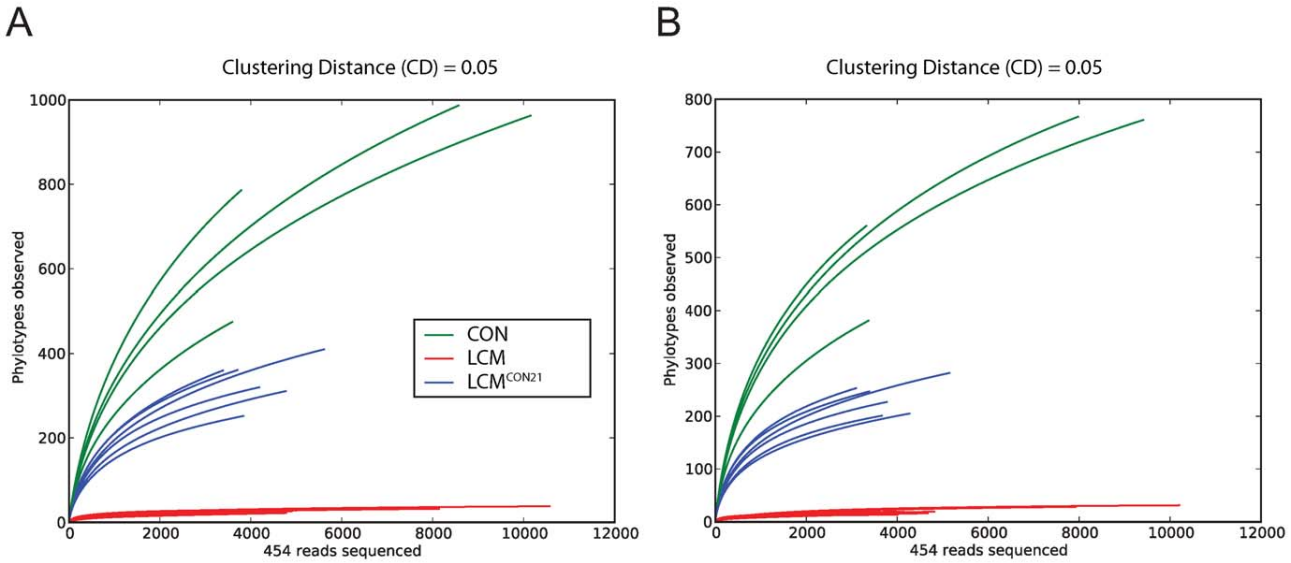


Figure 3. Collectors' curves of LCM, LCM^{CON21} and CON mice reveal different complexity. Collectors' curves were created for CD = 0.05 for each mouse from the total number of filtered sequences (A) or from chimera-removed sequences (B). *CON mice (green), LCM mice (red) and LCM^{CON21} mice (blue).*
doi:10.1371/journal.ppat.1000711.g003

Given the clear under-sampling and the sequencing strategy applied here, a species-level analysis is not conclusive and we decided to focus our further analysis at a higher taxonomic level (from the Family up to the Phylum).

Analysis of differences between CON and LCM^{CON21} gut microbiota

We next analyzed qualitative changes in microbiota composition during re-association. In particular, we focused at identifying which OTUs were transferred from the CON donor mice to the LCM recipients within 21 days. Those bacteria may contribute to protection against *S.* Typhimurium colonization.

In order to predict taxonomy for each OTU, we used either the reference sequence taxonomy information present within an OTU-cluster, if any, or the reads taxonomy predicted by the RDP classifier. To test if the taxonomy assignment via reference sequences provided a more resolved taxonomy, we compared taxonomy resolution obtained via reference sequences and via RDP-classifier annotated reads for OTUs which contained both reads and reference sequences (Fig. S7). For different CD and different taxon levels, the reference taxonomy always provided better taxonomic resolution from the phylum level (taxon_1) down to the genus level (taxon_5).

Euclidean distances between relative abundance profiles were computed for each mouse and every time-point sampled. Hierarchical clustering (average method) of all mice for taxon_2 (class) taxon_3 (order) and taxon_4 (family) were visualized on distinct heatmaps (Fig. 4; Fig. S8A,B). All CON mice (day 0 and day 21) clustered together as well as all the LCM mice before re-association. Additionally, we included two unmanipulated CON mice (donor 9855 and 9856) that were only sampled at one time-point to provide more samples of independent CON mice from the same mouse colony (n = 4 in total).

All samples of LCM mice from day 0 (before re-association) were highly similar and clustered together. The highest identity (determined by BLAST, all against all) between the V5V6 regions of the 8 different ASF members is of 93% (data not shown); therefore it is theoretically possible, for a small clustering distance, to detect each ASF species by our sequencing and taxonomy inference approach. Seven OTUs were systematically detected in the LCM mice, all assigned to the Firmicutes and Bacteroidetes phyla. Thus, we assume that the most abundant species in the feces of LCM-mice are ASF500 (Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; unclassified_Lachnospiraceae) and ASF519 (Bacteroidetes; Bacteroidetes; Bacteroidales; Porphyromonadaceae; Parabacteroides;). Abundance of ASF strains in different mice can be influenced by various factors [26,40]. Hence,

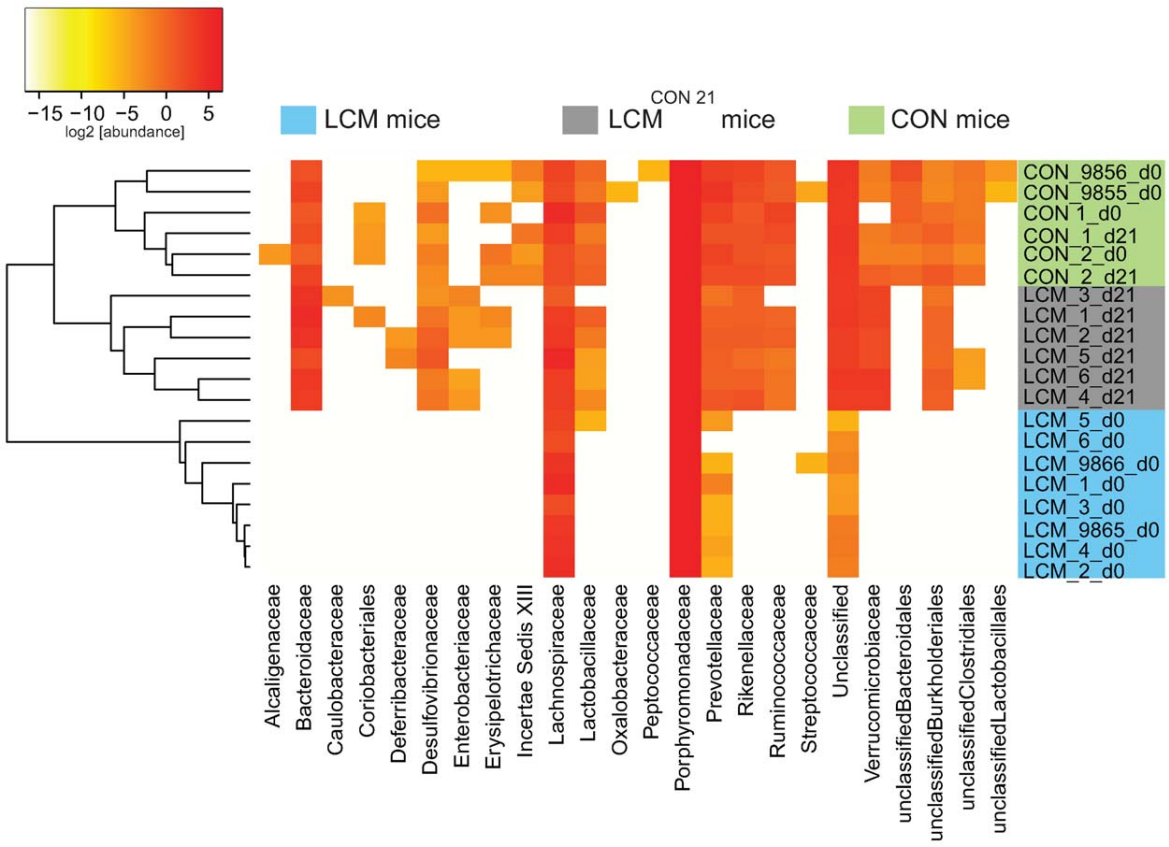


Figure 4. Heatmap showing OTU's distribution in different groups. Fecal microbiota of unmanipulated LCM mice was analyzed at day 0 (n = 8). 6 of these LCM mice (LCM_1 to LCM_6; blue) were conventionalized in two groups with 2 different CON-donors (CON_1 and CON_2; green) and fecal microbiota analyzed at day 21 (LCM_x_d21; grey). OTUs (CD = 0.05) were sorted according to taxon_4 (Family level; x-axis) and average clustering was performed on Euclidean distances calculated between abundance profiles for each mouse and every time-point sampled. Red color indicates high abundance (Log_2), yellow color low abundance. CON_9855_d0 and CON_9856_d0 and LCM_9865_d0 and LCM_9866_d0 are 2 additional CON or LCM mice, respectively sampled only at day 0.
doi:10.1371/journal.ppat.1000711.g004

the sampling depth could explain the non-detection of the other ASF members, which were most probably less abundant.

Gamma-Proteobacteria as indicators of susceptibility and resistance to *Salmonella*-infection

The qualitative microbiota analysis revealed that within the 21 days of re-association, bacteria from all detected phyla in the CON donor mice were transferred (Fig. 4; Fig. S8A,B). However, the gut microbiota of LCM^{CON21} was significantly less complex than that of CON mice, suggesting that the microbiota might also differ on a qualitative basis. This might be causally linked to the increased susceptibility to *Salmonella* infection. Thus, we compared the microbiota of CON and LCM^{CON21} with respect to lack or enrichment of specific clusters of bacteria (i.e. on order or family level). We analyzed which OTUs were significantly over- or underrepresented comparing LCM^{CON21} and CON mice. Interestingly, among others, OTUs assigned to the family of the Enterobacteriaceae were enriched in LCM^{CON21} mice, as compared to CON mice (Fig. 4; Dataset S1). Since *Salmonella* Typhimurium is also a member of the Enterobacteriaceae, the enrichment of such close relatives in LCM^{CON21} mice might be an indicator of favorable growth conditions for this type of bacteria.

This finding prompted us to investigate, whether there is a positive correlation between the abundance of Enterobacteriaceae (i.e. *E. coli*) and the susceptibility to *Salmonella* infection. We have previously observed that C57Bl/6 mice obtained from different sources (commercial breeders, other laboratories) exhibit differential degrees of CR against *Salmonella*. To analyze whether CR is linked to different *E. coli* titres, we defined fecal *E. coli* levels of mice from five different breedings (C57Bl6 background from our animal facility and others) before infecting them with *S.* Enteritidis wild type by oral gavage (5×10^7 cfu; no antibiotic-treatment). We used *S.* Enteritidis because pilot experiments in our laboratory had shown that this serovar generally leads to a higher disease incidence (colitis at day 4 after oral infection) in non-antibiotic-treated mice, than *S.* Typhimurium. *E. coli* is readily differentiated from other Enterobacteriaceae by colony color and morphology on MacConkey agar (see Materials and Methods for typing details). One day after infection, we determined fecal *S.* Enteritidis titers by plating. The mice were sacrificed at day 4 postinfection and we analyzed *S.* Enteritidis titers at systemic sites, in the intestine as well as cecal pathology (Fig. 5A; Fig. S9). Indeed, we observed a positive linear correlation between fecal *E. coli* levels before infection, *S.* Enteritidis colonization efficiency ($r^2 = 0.434$: Spearman $p = 0.0015$). If *S.* Enteritidis titres were above 1.5×10^5 cfu/g feces at day 1 p.i., mice developed colitis at day 4 p.i. This suggests that *E. coli* titres may predict whether mice are susceptible to *Salmonella* induced gut inflammation.

Higher levels of Lactobacilli predict higher intestinal colonization with a commensal *L. reuteri* ^{RR} after oral inoculation

We observed that higher *E. coli* levels positively correlate with increased *Salmonella* infectivity. This might be due to the close relatedness of these two species as they might have similar environmental requirements. Thus, we hypothesized that the same principle might apply for other intestinal bacteria. We tested this hypothesis using *Lactobacillus reuteri* ^{RR}, a rifampicin-resistant isolate from our mouse colony that can be specifically detected by culture [12].

We determined whether higher titres of intestinal Lactobacilli would correlate with increased gut colonization by *Lactobacillus reuteri* ^{RR} upon oral gavage. Lactobacilli are Gram-positive, of low G+C content, non-spore-forming, aerotolerant anaerobes and can be differentiated on selective media (i.e. MRS-agar). We determined fecal levels of Lactobacilli of mice from different sources and subsequently infected them with *Lactobacillus reuteri* ^{RR} (10^7 cfu by oral gavage). 1 and 5 days post infection we determined *Lactobacillus reuteri* ^{RR} titres in the feces. Indeed, we found significantly enhanced colonization of *Lactobacillus reuteri* ^{RR} in mice with higher titres of *Lactobacilli* (Fig. 5B). This suggests that, like in the case of *E. coli* and *Salmonella*, higher levels of Lactobacilli correlate with increased colonization efficiency by a commensal *Lactobacillus* strain.

Closely related phylotypes generally display significantly correlated abundances in the intestine

In order to investigate whether our observations with Enterobacteriaceae and Lactobacillaceae correspond to a more universal phenomenon that applies to closely related bacterial groups in general, we performed a systematic abundance correlation analysis between OTUs detected in 9 distinct CON mice (Fig. 6; Fig. S10). We limited our analysis to OTUs detected in at least 6 mice in order to lower the under-sampling bias in our 454 sequence data. Upon examination of OTUs defined at various CDs, we found that closely related phylotypes (i.e. $0<\text{reads divergence}<0.2$) generally display significantly correlated abundances (co-occurrence), more so than distantly related phylotypes. In summary, our results indicate that the invasion-success of novel species into a complex gut microbiota might be predetermined by the presence of closely related species or by factors that also influence the abundance of closely related species in this ecosystem.

Discussion

LCM mice as model for investigating the mechanisms of CR

It has been known for a long time, that the normal gut microbiota plays a key role in protection from infection with pathogenic bacteria. Germfree mice lack CR and thus are highly susceptible to infections with various pathogens [41]. They regain CR upon conventionalization with a normal microbiota [42]. This process has been studied extensively in the 1970s and 1980s; these earlier studies mainly addressed the question, which parts of the complex gut microbiota play a role in 'conventionalization' and inhibition of pathogen growth [43,44,45]. Due to technical limitations at that time, the studies were confined to the analysis of cultivated bacteria.

In contrast to this earlier work, we use LCM mice that are colonized with a stable, low-complexity gut microbiota being composed of typical gut bacteria i.e. *Bacteroides* spp., *Clostridium* spp., *Mucispirillum* spp. and Lactobacilli (ASF361, ASF457, ASF500 and ASF519) as starting point for the re-association studies. We show that LCM mice, despite being colonized with a numerically dense gut microbiota, are still susceptible to *Salmonella* gut infection and colitis. This system represents major advantages over the use of germfree mice. First of all, the maintenance of LCM mice is by far less extensive than that of germfree mice. We have maintained a colony of LCM mice for 24 months in IVC cages without significantly altering complexity of their gut microbiota. Therefore, LCM mice harbor 'typical' gut bacteria which, to some extent, protects against contamination with environmental bacteria, meaning that their gut ecosystem is somewhat normalized. Secondly, compared to germfree mice, the gut mucosal immune system and innate defense is partially normalized. Consequently, *Salmonella*-induced intestinal pathology is milder in LCM that in germfree mice (this work and [11]). However, the LCM gut microbiota apparently lacks certain parts

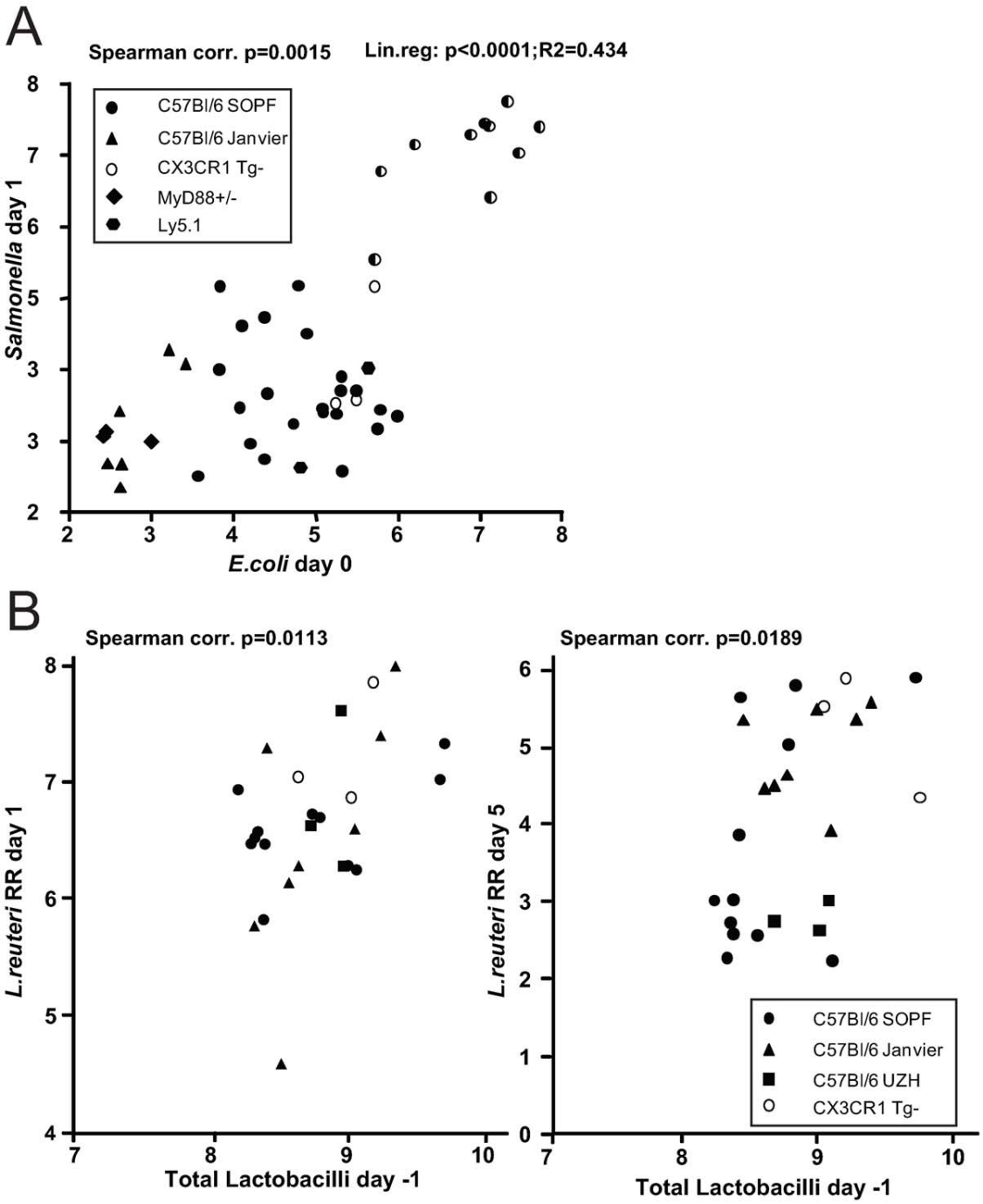


Figure 5. Infection experiments in conventional mice reveal correlation of bacterial infectivity with the prevalence of related species. (A) Groups normal unmanipulated CON mice (6-12 weeks; symbols indicate different sources) were infected with 5×10^7 cfu *S.* Enteritidis wild type by gavage. Fecal *E. coli* titres before infection were determined (x-axis; Log_{10} cfu/g). 1 day post infection, *S.* Enteritidis titres in the feces were determined (y-axis; Log_{10} cfu/g). Spearman and linear correlation were calculated ($p = 0.0015$; $p<0.0001$). The degree of gut inflammation was determined in the infected mice. Half-filled symbols indicate mice with inflammation score ≥ 4. (B) Groups normal unmanipulated CON mice (6-12 weeks; symbols indicate different sources) were infected with 5×10^7 *Lactobacillus reuteri*^{RR} (rifampicin-resistant) by gavage. Fecal levels of Lactobacilli were determined on MRS agar and plotted against fecal *Lactobacillus reuteri* ^{RR} titers at day 1 (left) and 5 (right) postinfection.
doi:10.1371/journal.ppat.1000711.g005

of the conventional gut microbiota important for protection against infection with enteropathogens (i.e. *Salmonella*, *E. coli*). Thus, LCM mice represent a very useful system to screen for protective bacteria and characterize the mode of protection.

What mechanisms underlie protection against enteropathogens by the gut microbiota? Host factors induced by bacterial colonization could be one mediator of CR. The gut microbiota instructs and shapes the mucosal innate and adaptive immunity

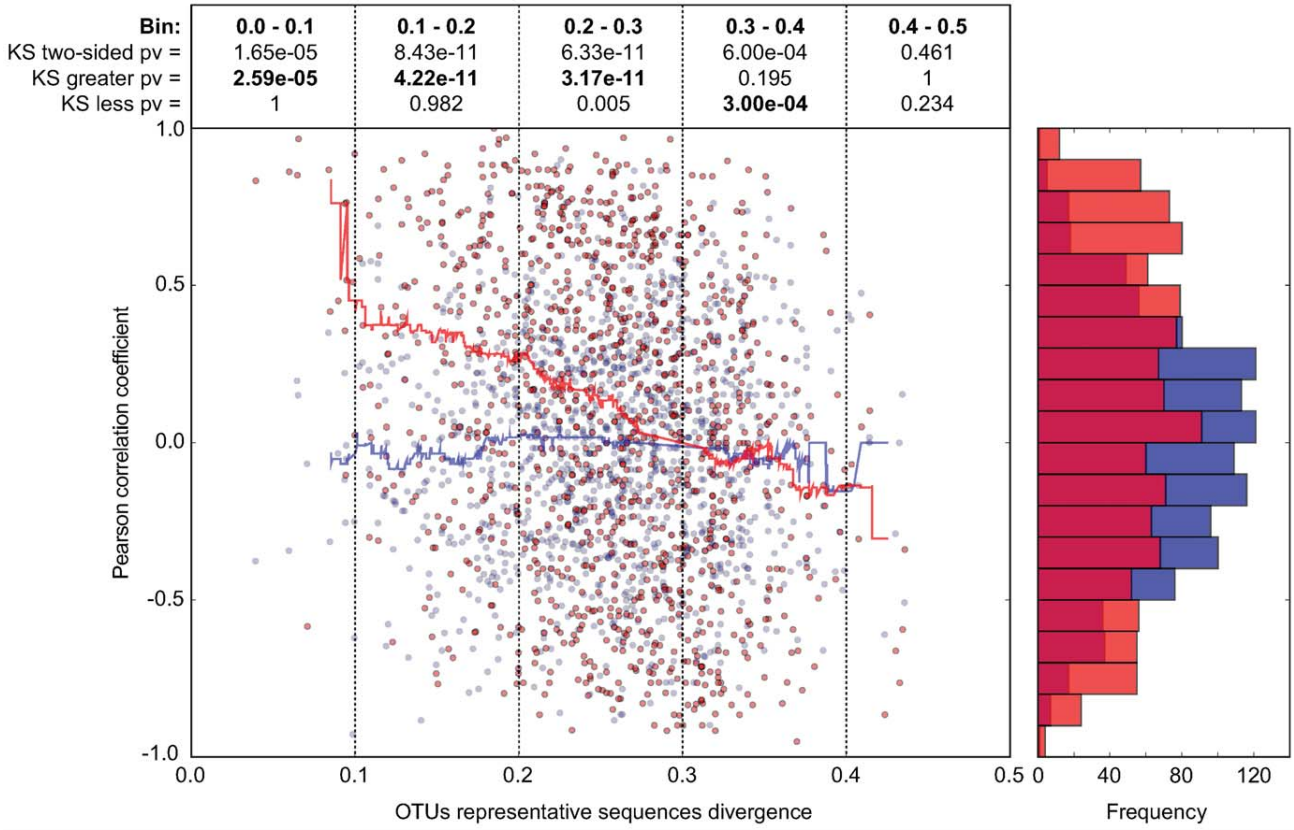


Figure 6. In CON mice, related bacterial lineages are preferentially observed together (quantitative co-occurrence). For all possible pairs of detected OTUs (i.e. present in at least 2/3 of the analyzed mice; CD = 0.2; each dot in the graph represents an OTU-pair), abundance correlations (y-axis, Pearson) were computed from abundance measurements in 9 distinct CON mice, and plotted against the molecular divergence between their representative 16S sequences (x-axis). The latter distances between representative sequences were computed using sequence identities as defined by the complete multiple alignment of all reads and all reference sequences. For hypothesis testing, we compared the data distribution (red) to a matched random distribution of OTU abundances generated by shuffling non-null OTU abundances between all OTUs (blue). Running medians are represented in the corresponding color. The Pearson correlation coefficient is −0.248 (p-value = 7.10e-18) for the actual data, and −0.017 (p-value = 0.563) for the randomized data. We compared the deviation of the actual data on the y-axis (Pearson correlation) from the distribution of the randomized data using the Kolmogorov-Smirnov test. Both two-sided and one-sided hypotheses (greater and less) were tested for each bin of 0.1 on the x-axis (0.0-0.1; 0.1–0.1; 0.2–0.3; 0.3–0.4; 0.4–0.5). Results are indicated in boxes in the upper part of the graph. Pv = P-value; 'KS two-sided pv' indicates whether there is a significant difference between the distribution of the data (red dots) and the distribution of the randomized data (blue dots); 'KS greater pv' indicates whether the Pearson correlation coefficient of the data (red dots) is significantly higher compared to the random background in a given bin (blue dots). 'KS less pv' indicates whether the Pearson correlation coefficient of the data (red dots) is significantly lower compared to the random background in (blue dots) in a given bin. The histogram on the right side of the graph represents the cumulative frequencies of the binned Pearson correlation coefficient data.
doi:10.1371/journal.ppat.1000711.g006

and keeps the host in a defense-competent state [46,47,48]. Alternatively the bacteria forming the gut ecosystem directly suppress pathogen growth. This could be mediated by blocking of pathogen receptor sites or the production of antibacterial substances and metabolic by-products like short-chain fatty acids (SCFA) [49,50,51,52,53]. Moreover, conventionalization also involves drastic changes in intestinal physiology, such as decrease of relative cecal size, free nutrient depletion, oxygen limitation and lowering of the redox potential [54,55,56]. The intestinal microbiota consists to the greatest part of obligate anaerobic and extremely oxygen-sensitive bacteria [57] and oxygen tension in the gut decreases gradually from stomach to rectum while bacterial density increases [58]. These conditions keep colonization levels of facultative aerobic bacteria, which comprise most enteropathogens, relatively low [57,59].

To date, the key bacteria inducing CR have not been unambiguously identified. Rolf Freter and coworkers aimed at identifying single strains that accomplish conventionalization of germfree mice. He demonstrated that a collection of 95 anaerobic intestinal isolates or even a combination of Clostridia and *Lactobacillus* spp. isolates is sufficient to restore CR [43,45]. To our knowledge, these 'CR-mediators' were never further described or characterized in detail. Since this would be a critical step towards understanding the molecular basis of CR, isolation and characterization of 'CR-mediators' will be subject of future analyses. To this end, LCM mice will be a useful tool.

Microbiota analysis during conventionalization

We analyzed microbiota changes during conventionalization of LCM mice using deep sequencing of 16S rRNA genes. This extends earlier studies that have focused only on culturable bacteria, to non-culturable strains. It is assumed that LCM mice pick up fecal bacteria from the CON donor mouse by coprophagy. The efficiency of microbiota transfer by coprophagy may be

questionable, since a great part of conventional gut microbiota is extremely oxygen sensitive. Still, we found that representatives of all five major eubacterial phyla typically present in the mammalian gut were transferred to LCM mice within 21 days. Although microbiota complexity drastically increased within 21 days of conventionalization, it was still significantly lower than in CON mice. Interestingly, in the same way as microbiota complexity, CR of LCM^{CON21} mice against *S.* Typhimurium was at a somewhat intermediate level. Still, LCM^{CON21} mice were, at least partially, protected from *Salmonella*-induced gut inflammation.

Overall complexity of LCM^{CON21} mice was not restored to the levels of CON mice, suggesting that conventionalization takes longer than 21 days to reach a high-density equilibrium state. For example, relatively few members of the Firmicutes and a high number of Verrucomicrobia were detected. Conventionalization is proposed to be a process of ecologic succession whereby the relative composition of the microbiota constantly changes, a sequence that mirrors microbiota-colonization after birth [60]. Alternatively, as the Firmicutes branch comprises most of the extremely oxygen-sensitive species, it is conceivable that they might be transferred less efficiently, or, only after oxygen tension in the gut is low enough to allow growth. It would be very interesting to analyze microbiota composition in fecal samples of LCM mice at different time points during conventionalization and also extend the analysis to longer time points beyond 21 days.

Levels of close relatives predict bacterial infectivity

Since LCM^{CON21} mice have partially gained CR during the 21 days re-association period, we aimed at identifying certain protective bacterial species, which are absent in LCM mice. Although microbiota complexity in LCM^{CON21} mice was too high for unequivocal species identification, comparative microbiota analysis detected an enrichment of Enterobacteriaceae in LCM^{CON21} mice. We concluded that this group of bacteria does not mediate CR ('CR-mediator') but may rather indicate the level of CR ('CR-indicator'). Upon screening a variety of conventional mice from different sources, we observed that higher *E. coli* titers positively correlated with *Salmonella* infectivity. Thus, *E. coli* can be regarded as 'CR-indicators' for *Salmonella* infection. Higher concentrations or diversity of Enterobactericeae can be indicative for alleviated CR [61,62]. This may explain why infection with 'CR-indicator' *E. coli* strains has been previously used as a method to judge the intensity of CR [63]. *E. coli* and *Salmonella* spp. are very close phylogenetic relatives. Strikingly, *E. coli* levels also correlated with susceptibility to *Salmonella*-induced colitis. When *E. coli* and concomitant *S.* Typhimurium levels were above 10^6 cfu/g at day 1 post pathogen infection, mice reliably developed gut inflammation. Interestingly, we found a similar correlation between the level of intrinsic Lactobacilli and the colonization levels of orally inoculated *Lactobacillus reuteri*^{RR} strain. Therefore, we speculated that the finding could be a general principle that applies to closely related bacterial groups in the intestinal ecosystem.

How can closely related species actually coexist in the same ecosystem? In theory, closely related species could occupy the same niche in the intestine although they had similar nutrient requirements or share the same adhesion receptors. However, in praxis, species A will perform slightly better than species B, which would lead to out-competition and elimination of B. Alternatively, species B could switch to the use of a different available nutrient source (or receptor) and coexist with species A in the same ecosystem. This principle has been demonstrated in case of *E. coli*. Different commensal *E. coli* strains can coexist in the intestine by using different nutrients [64].

But how is colonization level of a certain species A connected to the colonization efficiency of its close relative B? This might be explained by the fact that the same global selective pressure acts on both species. This global pressure could be the presence of a third species C that inhibits both A and B (i.e. by inhibitor production). Alternatively, A and B might have the same requirements of oxygen or the same sensitivity to antimicrobial peptides that only allows the bacteria to grow at a certain, defined density. This correlation would only be maintained, if none of the two strains produced a direct inhibitor against the other species (i.e. colicin, nisin, metabolites). Taken together, this principle suggested for Enterobacteriaceae and Lactobacillaceae might also apply for other bacterial groups, sharing common growth requirements.

General implications

Our data suggest, that subtle fluctuations in intestinal ecosystem composition between individuals might partly explain their differential susceptibility to gut infections or probiotic therapy. This knowledge could be exploited for screens of the human population to identify certain risk-or susceptibility groups. This would then enable the correlation of these data to other parameters (lifestyle, age, gender, nutrition). The existence of a highly dynamic niche for growth of Enterobacteriaceae, varying between different individuals, might reflect the differential susceptibility to gut infections within the human population. Some patients might have suffered from insults that induce a transient 'out of equilibrium' state of the microbiota that renders it less protective. Such conditions could be nutrient deficiencies, stress, illness or a history of antibiotic treatment. Screening of people at risk (elderly, immune-suppressed) might thus help in early disease prevention and potentially enable more targeted use of antibiotics.

Supporting Information

Figure S1 LCM and smCON mice develop inflammation of the cecum and colon after *S.* Tm infection. HE-stained tissue cross sections (see M&M) of the cecum, proximal and distal colon of (A) a naïve CON, (B), a smCON mouse at day 3 post infection with *S.* Tm wild type and (C), LCM mouse at day 3 post infection with *S.* Tm wild type. Enlarged section (black box) is shown in the right panels. Scale bar: 50 μm.
Found at: doi:10.1371/journal.ppat.1000711.s001 (3.21 MB PDF)

Figure S2 Avirulent *S.* Typhimurium do not induce inflammation in LCM mice. Groups (n = 5) of LCM mice were infected for 3 days with *S.* Typhimurium wild type or the avirulent mutant *S.* Typhimurium^{avir} (Δ*invG*; *sseD*::*aphT*). *S.* Typhimurium levels in the feces at day 1 post infection (A), cecal content (B), mLN (C), spleen (D). (E) Cecal pathology scored in HE-stained tissue sections (see M&M). (F) HE-stained sections of cecal tissue from indicated mice. Enlarged section (white box) is shown in the lower panel. Scale bar: 100 μm.
Found at: doi:10.1371/journal.ppat.1000711.s002 (0.50 MB PDF)

Figure S3 Microbiota density in CON and LCM mice. (A) Cecal content of CON and LCM mice was stained with Sytox-green and bacteria were counted in a Neubauer-chamber. Bacterial density is given as Sytox-green positive bacteria per gram cecal content. (B) Representative confocal fluorescence microscopy images of cecum tissue sections from the mice shown in (A). Nuclei and bacterial DNA are stained by Sytox-green (green), the epithelial brush border actin by phalloidin-Alexa-647 (blue). Scale bar: 50μm.
Found at: doi:10.1371/journal.ppat.1000711.s003 (0.49 MB PDF)

Figure S4 Example for a chimeric sequence read obtained by pyrosequencing. The figure depicts an example for a chimeric V5-V6 read. The mismatches of the chimeric read to the best BLASTn hits (best bit-score) using either a non-redundant (nr) Bacteroidetes database (a) or a nr Firmicutes database (b). Using our chimera detection approach, the V5 region of the depicted read was predicted as Firmicutes and its V6 region as Bacteroidetes.
Found at: doi:10.1371/journal.ppat.1000711.s004 (0.08 MB PDF)

Figure S5 Detection of ASF bacteria in LCM mice by PCR. Fecal bacterial DNA was extracted from LCM mice and used as template for ASF-strain specific PCR. Strain specific PCR for ASF356 (417 bp), ASF360 (131 bp), ASF361 (182 bp), ASF457 (95 bp), ASF492 (167 bp), ASF500 (285 bp), ASF502 (427 bp), and ASF519 (429 bp) (expected sizes shown in parentheses) was performed as described [26]. Linearized plasmids as positive control, containing the 16SrRNA gene of each ASF strain (A) or fecal bacterial DNA was used as template (B). Only strains ASF361, ASF457, ASF500 and ASF519 could be detected in the feces of LCM mice (here only 1 representative PCR result shown).
Found at: doi:10.1371/journal.ppat.1000711.s005 (0.02 MB PDF)

Figure S6 Collectors' curves of LCM, LCM^{CON21} and CON mice reveal different complexity. Collectors' curves were created for different CD (0.1; 0.03; 0.01) for each mouse from the total number of filtered sequences (A) or from chimera-removed sequences (B). CON mice (green), LCM mice (red) and LCM^{CON21} mice (blue).
Found at: doi:10.1371/journal.ppat.1000711.s006 (0.68 MB PDF)

Figure S7 Taxonomy resolution obtained via reference sequences and via RDP-classifier. OTU taxonomy assignment was inferred preferentially using reference sequences annotations (if present in the cluster) or using RDP classifier predictions on the 454 reads. For OTUs containing both reference sequences and reads, we compared the resolution of taxonomies assigned by majority vote via both approaches. For each clustering distance tested and for all taxon levels, a more resolved taxonomy was obtained using the reference sequences annotations.
Found at: doi:10.1371/journal.ppat.1000711.s007 (0.10 MB PDF)

Figure S8 Heatmap showing OTU's distribution in different groups at different phylogenic resolutions. Analysis of fecal microbiota of the mice shown in Fig. 4. Fecal microbiota of unmanipulated LCM mice was analyzed at day 0 (n = 8). 6 of these LCM mice (LCM_1 to LCM_6; blue) were conventionalized in two groups with 2 different CON-donors (CON_1 and CON_2; green) and fecal microbiota analyzed at day 21 (LCM_x_d21; grey). (A) OTUs were sorted according to taxon_2 (class level; X-axis) of (B) according to taxon_3 (order level; X-axix) and average clustering was performed on Euclidean distances calculated between abundance profiles for each mouse and every time-point sampled. Red color indicates high abundance (Log_2), yellow color low abundance. CON_9855_d0 and CON_9856_d0 and LCM_9865_d0 and LCM_9866_d0are 2 additional CON or LCM mice, respectively sampled only at day 0.
Found at: doi:10.1371/journal.ppat.1000711.s008 (0.15 MB PDF)

Figure S9 Development of *Salmonella*-induced gut inflammation in mice with differential fecal *E. coli* titres. Groups of normal unmanipulated CON mice (6–12 weeks; symbols indicate different sources; x-axis) were infected with 5×10^7 cfu *S.* Enteritidis wild type by gavage and sacrificed at day 4 p.i. (see Fig. 5A). Cecal pathology of mice was analyzed in HE-stained tissue sections (see Materials and Methods).
Found at: doi:10.1371/journal.ppat.1000711.s009 (0.14 MB PDF)

Figure S10 In CON mice, related bacterial lineages are preferentially observed together (quantitative co-occurrence). For all possible pairs of detected OTUs (i.e. present in 2/3 of the analyzed mice; A, CD = 0.1 and B, CD = 0.05; each dot in the graph represents an OTU-pair), abundance correlations (y-axis, Pearson) were computed from abundance measurements in 9 distinct CON mice, and plotted against the molecular divergence between their representative 16S sequences (x-axis). Running medians are represented in the corresponding color. We compared the deviation of the actual data on the y-axis (Pearson correlation) from the distribution of the randomized data using the Kolmogorov-Smirnov test. Both two-sided and one-sided hypotheses (greater and less) were tested for each bin of 0.1 on the x-axis (0.0–0.1; 0.1–0.1; 0.2–0.3; 0.3–0.4). Results are indicated in boxes in the upper part of each graph. Pv = P-value; 'KS two-sided pv' indicates whether there is a significant difference between the distribution of the data (red dots) and the distribution of the randomized data (blue dots); 'KS greater pv' indicates whether the Pearson correlation coefficient of the data (red dots) is significantly higher compared to the random background in (blue dots) in a given bin. 'KS less pv' indicates whether the Pearson correlation coefficient of the data (red dots) is significantly lower compared to the random background in (blue dots) in a given bin. The histogram on the right side of the graph represents the cumulative frequencies of the binned Pearson correlation coefficient data.
Found at: doi:10.1371/journal.ppat.1000711.s010 (6.78 MB PDF)

Table S1 Parameters of microbial complexity of CON-donors day 0 (n = 4).
Found at: doi:10.1371/journal.ppat.1000711.s011 (0.06 MB DOC)

Table S2 Parameters of microbial complexity of LCM-recipients day 0 (n = 8).
Found at: doi:10.1371/journal.ppat.1000711.s012 (0.03 MB DOC)

Table S3 Parameters of microbial complexity of LCM-recipients day 21 (n = 6).
Found at: doi:10.1371/journal.ppat.1000711.s013 (0.03 MB DOC)

Dataset S1 Comparative OTU abundance analysis of LCFd21 and CON mice showing differentially abundant OTU's at different Clustering Distances (ClustD).
Found at: doi:10.1371/journal.ppat.1000711.s014 (0.07 MB XLS)

Author Contributions

Conceived and designed the experiments: BS SC RK CvM AJM WDH. Performed the experiments: BS SC RK SH MS. Analyzed the data: BS SC RK. Contributed reagents/materials/analysis tools: SF TCW JK KDM AJM. Wrote the paper: BS SC WDH. Applied for funding: BS CvM AJM WDH.

References

1. Palmer C, Bik EM, Digiulio DB, Relman DA, Brown PO (2007) Development of the Human Infant Intestinal Microbiota. PLoS Biol 5: e177. doi:10.1371/journal.pbio.0050177.

2. Dethlefsen L, Huse S, Sogin ML, Relman DA (2008) The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing. PLoS Biol 6: e280. doi:10.1371/journal.pbio.0060280.

3. Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, et al. (2008) Evolution of mammals and their gut microbes. Science 320: 1647–1651.
4. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2009) A core gut microbiome in obese and lean twins. Nature 457: 480–484.
5. van der Waaij D, Berghuis-de Vries JM, Lekkerkerk L-v (1971) Colonization resistance of the digestive tract in conventional and antibiotic-treated mice. J Hyg (Lond) 69: 405–411.
6. Wilson KH, Sheagren JN, Freter R, Weatherbee L, Lyerly D (1986) Gnotobiotic models for study of the microbial ecology of Clostridium difficile and Escherichia coli. J Infect Dis 153: 547–551.
7. Stecher B, Hardt WD (2008) The role of microbiota in infectious disease. Trends Microbiol 16: 107–114.
8. Que JU, Hentges DJ (1985) Effect of streptomycin administration on colonization resistance to Salmonella typhimurium in mice. Infect Immun 48: 169–174.
9. Kelly CP, LaMont JT (1998) Clostridium difficile infection. Annu Rev Med 49: 375–390.
10. Vollaard EJ, Clasener HA, van Saene HK, Muller NF (1990) Effect on colonization resistance: an important criterion in selecting antibiotics. Dicp 24: 60–66.
11. Stecher B, Macpherson AJ, Hapfelmeier S, Kremer M, Stallmach T, et al. (2005) Comparison of Salmonella enterica Serovar Typhimurium Colitis in Germfree Mice and Mice Pretreated with Streptomycin. Infect Immun 73: 3228–3241.
12. Stecher B, Robbiani R, Walker AW, Westendorf AM, Barthel M, et al. (2007) Salmonella enterica Serovar Typhimurium Exploits Inflammation to Compete with the Intestinal Microbiota. PLoS Biol 5: e244. doi:10.1371/journal.-pbio.0050244.
13. Saxelin M, Pessi T, Salminen S (1995) Fecal recovery following oral administration of Lactobacillus strain GG (ATCC 53103) in gelatine capsules to healthy volunteers. Int J Food Microbiol 25: 199–203.
14. Alander M, Satokari R, Korpela R, Saxelin M, Vilpponen-Salmela T, et al. (1999) Persistence of colonization of human colonic mucosa by a probiotic strain, Lactobacillus rhamnosus GG, after oral consumption. Appl Environ Microbiol 65: 351–354.
15. Dunne C, Murphy L, Flynn S, O'Mahony L, O'Halloran S, et al. (1999) Probiotics: from myth to reality. Demonstration of functionality in animal models of disease and in human clinical trials. Antonie Van Leeuwenhoek 76: 279–292.
16. Prilassnig M, Wenisch C, Daxboeck F, Feierl G (2007) Are probiotics detectable in human feces after oral uptake by healthy volunteers? Wien Klin Wochenschr 119: 456–462.
17. Denou E, Pridmore RD, Berger B, Panoff JM, Arigoni F, et al. (2008) Identification of genes associated with the long-gut-persistence phenotype of the probiotic Lactobacillus johnsonii strain NCC533 using a combination of genomics and transcriptome analysis. J Bacteriol 190: 3161–3168.
18. Barthel M, Hapfelmeier S, Quintanilla-Martinez L, Kremer M, Rohde M, et al. (2003) Pretreatment of mice with streptomycin provides a Salmonella enterica serovar Typhimurium colitis model that allows analysis of both pathogen and host. Infect Immun 71: 2839–2858.
19. Jung S, Aliberti J, Graemmel P, Sunshine MJ, Kreutzberg GW, et al. (2000) Analysis of fractalkine receptor CX(3)CR1 function by targeted deletion and green fluorescent protein reporter gene insertion. Mol Cell Biol 20: 4106–4114.
20. Adachi O, Kawai T, Takeda K, Matsumoto M, Tsutsui H, et al. (1998) Targeted disruption of the MyD88 gene results in loss of IL-1- and IL-18-mediated function. Immunity 9: 143–150.
21. Hapfelmeier S, Muller AJ, Stecher B, Kaiser P, Barthel M, et al. (2008) Microbe sampling by mucosal dendritic cells is a discrete, MyD88-independent step in DeltainvG S. Typhimurium colitis. J Exp Med 205: 437–450.
22. Suar M, Jantsch J, Hapfelmeier S, Kremer M, Stallmach T, et al. (2006) Virulence of broad- and narrow-host-range Salmonella enterica serovars in the streptomycin-pretreated mouse model. Infect Immun 74: 632–644.
23. Hoiseth SK, Stocker BA (1981) Aromatic-dependent *Salmonella typhimurium* are non-virulent and effective as live vaccines. Nature 291: 238–239.
24. Hapfelmeier S, Ehrbar K, Stecher B, Barthel M, Kremer M, et al. (2004) Role of the Salmonella Pathogenicity Island 1 Effector Proteins SipA, SopB, SopE, and SopE2 in Salmonella enterica Subspecies 1 Serovar Typhimurium Colitis in Streptomycin-Pretreated Mice. Infect Immun 72: 795–809.
25. White JR, Nagarajan N, Pop M (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol 5: e1000352. doi:10.1371/journal.pcbi.1000352.
26. Sarma-Rupavtarm RB, Ge Z, Schauer DB, Fox JG, Polz MF (2004) Spatial distribution and stability of the eight microbial species of the altered schaedler flora in the mouse gastrointestinal tract. Appl Environ Microbiol 70: 2791–2800.
27. Weisburg WG, Barns SM, Pelletier DA, Lane DJ (1991) 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol 173: 697–703.
28. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37: D141–145.
29. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2009) Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates. Environ Microbiol.
30. Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, et al. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet 4: e1000255. doi:10.1371/journal.pgen.1000255.
31. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72: 5069–5072.
32. Eddy SR (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics 3: 18.
33. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73: 5261–5267.
34. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A 103: 12115–12120.
35. Kanagawa T (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). J Biosci Bioeng 96: 317–323.
36. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, et al. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods 6: 639–641.
37. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2006) New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. Appl Environ Microbiol 72: 5734–5741.
38. Sekirov I, Tam NM, Jogova M, Robertson ML, Li Y, et al. (2008) Antibiotic-induced perturbations of the intestinal microbiota alter host susceptibility to enteric infection. Infect Immun 76: 4726–4736.
39. Dewhirst FE, Chien CC, Paster BJ, Ericson RL, Orcutt RP, et al. (1999) Phylogeny of the defined murine microbiota: altered Schaedler flora. Appl Environ Microbiol 65: 3287–3292.
40. Ge Z, Feng Y, Taylor NS, Ohtani M, Polz MF, et al. (2006) Colonization dynamics of altered Schaedler flora is influenced by gender, aging, and Helicobacter hepaticus infection in the intestines of Swiss Webster mice. Appl Environ Microbiol 72: 5100–5103.
41. Collins FM, Carter PB (1978) Growth of salmonellae in orally infected germfree mice. Infect Immun 21: 41–47.
42. Koopman JP, Kennis HM, Mullink JW, Prins RA, Stadhouders AM, et al. (1984) 'Normalization' of germfree mice with anaerobically cultured caecal flora of 'normal' mice. Lab Anim 18: 188–194.
43. Freter R, Abrams GD (1972) Function of various intestinal bacteria in converting germfree mice to the normal state. Infect Immun 6: 119–126.
44. Freter R, Brickner H, Botney M, Cleven D, Aranki A (1983) Mechanisms that control bacterial populations in continuous-flow culture models of mouse large intestinal flora. Infect Immun 39: 676–685.
45. Itoh K, Freter R (1989) Control of Escherichia coli populations by a combination of indigenous clostridia and lactobacilli in gnotobiotic mice and continuous-flow cultures. Infect Immun 57: 559–565.
46. Smith K, McCoy KD, Macpherson AJ (2007) Use of axenic animals in studying the adaptation of mammals to their commensal intestinal microbiota. Semin Immunol 19: 59–69.
47. Cash HL, Whitham CV, Behrendt CL, Hooper LV (2006) Symbiotic bacteria direct expression of an intestinal bactericidal lectin. Science 313: 1126–1130.
48. Brandl K, Plitas G, Schnabl B, DeMatteo RP, Pamer EG (2007) MyD88-mediated signals induce the bactericidal lectin RegIII gamma and protect mice against intestinal Listeria monocytogenes infection. J Exp Med 204: 1891–1900.
49. Candela M, Perna F, Carnevali P, Vitali B, Ciati R, et al. (2008) Interaction of probiotic Lactobacillus and Bifidobacterium strains with human intestinal epithelial cells: adhesion properties, competition against enteropathogens and modulation of IL-8 production. Int J Food Microbiol 125: 286–292.
50. Filho-Lima JV, Vieira EC, Nicoli JR (2000) Antagonistic effect of Lactobacillus acidophilus, Saccharomyces boulardii and Escherichia coli combinations against experimental infections with Shigella flexneri and Salmonella enteritidis subsp. typhimurium in gnotobiotic mice. J Appl Microbiol 88: 365–370.
51. Cursino L, Smajs D, Smarda J, Nardi RM, Nicoli JR, et al. (2006) Exoproducts of the Escherichia coli strain H22 inhibiting some enteric pathogens both in vitro and in vivo. J Appl Microbiol 100: 821–829.
52. Millette M, Cornut G, Dupont C, Shareck F, Archambault D, et al. (2008) Capacity of human nisin- and pediocin-producing lactic Acid bacteria to reduce intestinal colonization by vancomycin-resistant enterococci. Appl Environ Microbiol 74: 1997–2003.
53. Gantois I, Ducatelle R, Pasmans F, Haesebrouck F, Hautefort I, et al. (2006) Butyrate specifically down-regulates salmonella pathogenicity island 1 gene expression. Appl Environ Microbiol 72: 946–949.
54. Koopman JP, Janssen FG, van Druten JA (1975) Oxidation-reduction potentials in the cecal contents of rats and mice. Proc Soc Exp Biol Med 149: 995–999.
55. Macfarlane GT, Macfarlane S (1997) Human colonic microbiota: ecology, physiology and metabolic potential of intestinal bacteria. Scand J Gastroenterol Suppl 222: 3–9.
56. Macfarlane S, Macfarlane GT (2003) Regulation of short-chain fatty acid production. Proc Nutr Soc 62: 67–72.
57. Savage DC (1977) Microbial ecology of the gastrointestinal tract. Annu Rev Microbiol 31: 107–133.
58. He G, Shankar RA, Chzhan M, Samouilov A, Kuppusamy P, et al. (1999) Noninvasive measurement of anatomic structure and intraluminal oxygenation

in the gastrointestinal tract of living mice with spatial and spectral EPR imaging. Proc Natl Acad Sci U S A 96: 4586–4591.
59. Komitopoulou E, Bainton NJ, Adams MR (2004) Premature Salmonella Typhimurium growth inhibition in competition with other Gram-negative organisms is redox potential regulated via RpoS induction. J Appl Microbiol 97: 964–972.
60. Lee A, Gemmell E (1972) Changes in the mouse intestinal microflora during weaning: role of volatile fatty acids. Infect Immun 5: 1–7.
61. Vollaard EJ, Clasener HA (1994) Colonization resistance. Antimicrob Agents Chemother 38: 409–414.
62. Apperloo-Renkema HZ, Van der Waaij BD, Van der Waaij D (1990) Determination of colonization resistance of the digestive tract by biotyping of Enterobacteriaceae. Epidemiol Infect 105: 355–361.
63. van der Waaij D (1983) Colonization pattern of the digestive tract by potentially pathogenic microorganisms: colonization-controlling mechanisms and consequences for antibiotic treatment. Infection 11 Suppl 2: S90–92.
64. Leatham MP, Banerjee S, Autieri SM, Mercado-Lubo R, Conway T, et al. (2009) Precolonized human commensal Escherichia coli strains serve as a barrier to E. coli O157:H7 growth in the streptomycin-treated mouse intestine. Infect Immun 77: 2876–2886.

5 INSIGHTS ON MICROBIAL PHYSIOLOGY BY ENVIRONMENTAL OMICS DATA INTEGRATION

5.1 PREFACE

Plants wear their gut on the outside. Indeed, the plant phyllosphere (i. e., the aerial parts of plants) is colonized by microorganisms (pathogens or commensals) that can influence their hosts on the level of the individual plant. The microbiota present on the roughly one billion square kilometers of world-wide leaf surfaces is sufficiently abundant to have a significant impact on the global carbon and nitrogen cycles on Earth. In the following study, we combined high-throughput DNA sequencing technology and MS-based proteomics to analyze entire bacterial phyllosphere communities on different plant species *in situ*. The analysis included the development of a new methodology to integrate both genomic and proteomic datasets and helped us to characterize the diversity and physiology of the phyllosphere colonizers.

For this work[22], I realized all the computational analyses dealing with the examination of 16S rRNA sequences (rarefaction analysis, figure 3) and the annotation of environmental genomics data (including the Pfam domains annotation in figure 4). I also developed a pipeline for the visualization and integration of environmental genomic and proteomic data in order to estimate gene expression levels from the most abundant taxa detected in the phyllosphere microbial communities (see supplemental figures S4 and S5).

5.2 COMMUNITY PROTEOGENOMICS

The publication is included below.

Community proteogenomics reveals insights into the physiology of phyllosphere bacteria

Nathanaël Delmotte[a,1], Claudia Knief[a,1], Samuel Chaffron[b], Gerd Innerebner[a], Bernd Roschitzki[c], Ralph Schlapbach[c], Christian von Mering[b], and Julia A. Vorholt[a,2]

[a]Institute of Microbiology, Eidgenössische Technische Hochschule Zurich, Wolfgang-Pauli-Strasse 10, 8093 Zurich, Switzerland; [b]Institute of Molecular Biology and Swiss Institute of Bioinformatics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland; and [c]Functional Genomics Center Zurich, University of Zurich/Eidgenössische Technische Hochschule Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

Edited by Steven E. Lindow, University of California, Berkeley, CA, and approved July 16, 2009 (received for review May 12, 2009)

Aerial plant surfaces represent the largest biological interface on Earth and provide essential services as sites of carbon dioxide fixation, molecular oxygen release, and primary biomass production. Rather than existing as axenic organisms, plants are colonized by microorganisms that affect both their health and growth. To gain insight into the physiology of phyllosphere bacteria under in situ conditions, we performed a culture-independent analysis of the microbiota associated with leaves of soybean, clover, and *Arabidopsis thaliana* plants using a metaproteogenomic approach. We found a high consistency of the communities on the 3 different plant species, both with respect to the predominant community members (including the alphaproteobacterial genera *Sphingomonas* and *Methylobacterium*) and with respect to their proteomes. Observed known proteins of *Methylobacterium* were to a large extent related to the ability of these bacteria to use methanol as a source of carbon and energy. A remarkably high expression of various TonB-dependent receptors was observed for *Sphingomonas*. Because these outer membrane proteins are involved in transport processes of various carbohydrates, a particularly large substrate utilization pattern for Sphingomonads can be assumed to occur in the phyllosphere. These adaptations at the genus level can be expected to contribute to the success and coexistence of these 2 taxa on plant leaves. We anticipate that our results will form the basis for the identification of unique traits of phyllosphere bacteria, and for uncovering previously unrecorded mechanisms of bacteria-plant and bacteria-bacteria relationships.

metaproteomics | methylotrophy | plant phyllosphere | *Pseudomonas* | *Sphingomonas*

For terrestrial plants, the phyllosphere represents the interface between the above-ground parts of plants and the air. Conservative estimates indicate that the roughly 1 billion square kilometers of worldwide leaf surfaces host more than 10^{26} bacteria, which are the most abundant colonizers of this habitat (1, 2). The overall microbiota in this ecosystem is thus sufficiently large to have an impact on the global carbon and nitrogen cycles. Additionally, the phyllosphere inhabitants influence their hosts at the level of the individual plants. To a large extent, interest in phyllosphere microbiology has been driven by investigations on plant pathogens. Their spread, colonization, survival, and pathogenicity mechanisms have been the subject of numerous studies (2). Much less understood are nonpathogenic microorganisms that inhabit the phyllosphere. The composition of the phyllosphere microbiota has been analyzed in only a few studies by cultivation-independent methods (e.g., refs. 3–5); however, such methods are essential in light of the yet uncultivated majority of bacteria existing in nature (6), or more specifically on plant leaves (7). Not only their identity, but in particular the physiological properties of phyllosphere bacteria, their adaptations to the habitat, and their potential role (e.g., with respect to modulating population sizes of pathogens) remain largely unknown. Current knowledge on the traits important in the phyllosphere is derived from relatively few studies on gene expression and stems mostly from model bacteria cultivated on host plants under controlled conditions (8–11). However, under natural conditions, plants and their residing microorganisms are exposed to a host of diverse, highly variable environmental factors, including UV light, temperature, and water availability; moreover, individual microbes are subjected to competition with other microorganisms over resources, such as nutrients and space.

Toward a deeper understanding of phyllosphere microbiology, and in particular to learn more about the commensal majority of plant leaf colonizing bacteria, which may be of relevance for plant health and development, integrated approaches are needed. Here, we combined metagenomic and metaproteomic approaches (community proteogenomics) (12) to analyze bacterial phyllosphere communities in situ (the phyllosphere is defined here as the environment comprising both the surface and the apoplast of leaves). We studied 3 different plant species grown under standard agriculture regimes or under natural conditions. Our results provide insight into the physiology of bacteria and point toward common adaptation mechanisms among the phyllosphere populations of different plants.

Results and Discussion

The prokaryotic phyllosphere populations in our study were obtained from 2 field-grown plant species, soybean (*Glycine max*, 2 samples) and clover (*Trifolium repens*, 3 samples), as well as from a wild population of the model plant *Arabidopsis thaliana* (1 sample) (Fig. 1, Table S1). Genomic DNA and proteins of the prokaryotes were extracted from the same pools of cells. For 1 of the 6 samples, Soybean 2, 260 Mbp of metagenomic sequence reads were generated using 454 pyrosequencing technology.

Microbial Community Composition. To characterize the composition of the phyllosphere microbiota, we applied complementary approaches: phylogenetic information was derived from protein-coding marker genes in the metagenome database generated in this study, as well as from 16S rRNA gene-based clone libraries. Comparative community analyses were additionally done by denaturing gradient gel electrophoresis (DGGE) to evaluate the representativeness of the samples.

Author contributions: N.D., C.K., and J.A.V. designed research; N.D., C.K., G.I., and J.A.V. performed research; S.C., B.R., R.S., and C.v.M. contributed new reagents/analytic tools; N.D., C.K., S.C., G.I., C.v.M., and J.A.V. analyzed data; and N.D., C.K., C.v.M., and J.A.V. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: Protein databases as well as lists of identified proteins in form of Scaffold and Excel files are available at http://www.micro.biol.ethz.ch/research/vorholt. MS/MS data have been deposited in the PRIDE database (9850–9860) and gene sequences in the GenBank database [accession nos. 38721 (Metagenome), and FN421480 to FN421999 (16S rRNA)].

[1]N.D. and C.K. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: vorholt@micro.biol.ethz.ch.

This article contains supporting information online at www.pnas.org/cgi/content/full/0905240106/DCSupplemental.

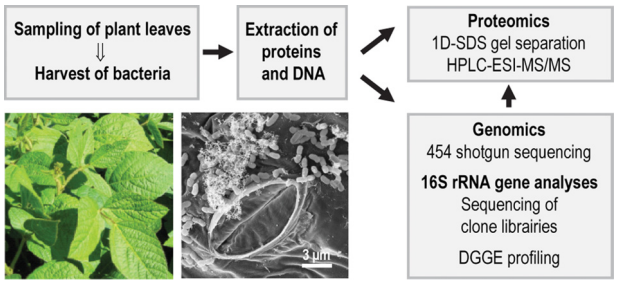


Fig. 1. Experimental strategy applied to characterize the phyllosphere microbiota. All analyses described were conducted from identical pools of cells as starting material. The photograph shows leaves of soybean plants; the electron micrograph shows the surface of an *Arabidopsis* leaf.

In a first step, the phylogenetic information contained in selected protein-coding marker genes of the metagenome data were used to analyze the composition of the microbial phyllosphere community in the Soybean 2 sample (Fig. 2). This approach gives a quantitative overview without the introduction of a PCR primer bias (13). Overall, we observed a clear dominance of Alphaproteobacteria. A relevant fraction of this group is well known to have adopted an extra- or intracellular lifestyle as plant mutualists or as plant or animal pathogens. The majority of Alphaproteobacteria in the Soybean 2 sample belonged to the families of Sphingomonadaceae (*Sphingomonas* 20.1%, *Novosphingobium* 10.1%) and Methylobacteriaceae (*Methylobacterium* 20.2%), which have been previously detected on plants (see, for example, refs. 14–16). Bacteria of the genus *Methylobacterium* and *Sphingomonas* were also detected in the Soybean 2 sample by 16S rRNA gene-based community anal-

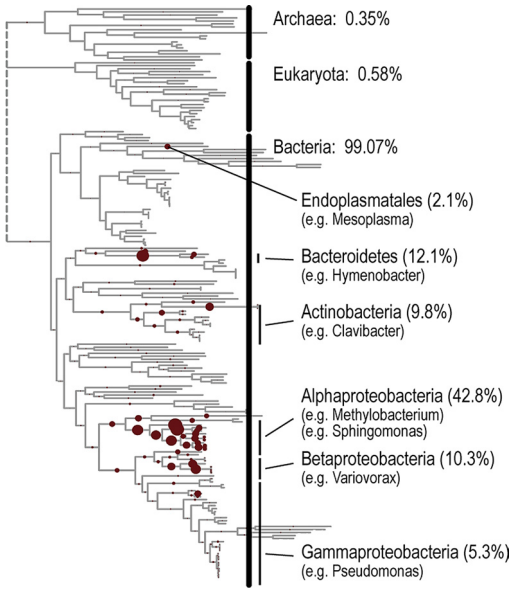


Fig. 2. Taxonomic composition of the bacterial community in the Soybean 2 sample. A phylogenetic tree calculated from informative marker genes of completely sequenced organisms serves as a reference onto which the estimated coverage of the most abundant clades present in the Soybean 2 sample is projected. Coverage is estimated based on the quantity of marker genes found in the metagenome data and is indicated by red dots (13). A selection of typical representatives of the clades is listed to the right, annotated according to the 16S rRNA gene-sequencing results (Table S2). Archaea contributed only 0.35% to the microbial community of the sample and were identified as members of the mesophilic Crenarchaeota (group 1.1b) by 16S rRNA gene sequencing. The low contribution of eukaryotes (0.58%) to the analyzed phyllosphere community in the soybean sample is in accordance with the design of the microbial harvesting procedure, which included a physical depletion step for eukaryotic cells.

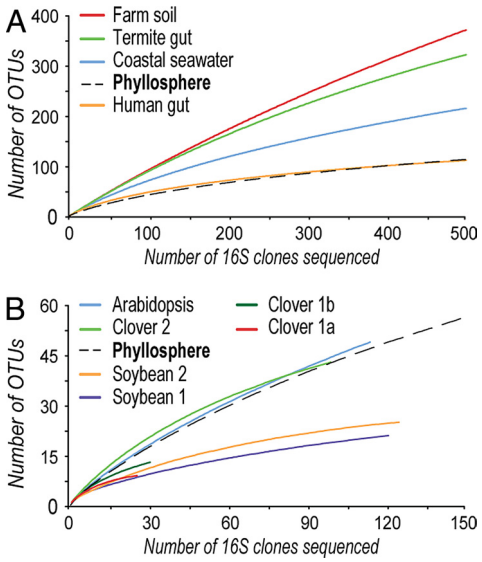


Fig. 3. Rarefaction analysis of 16S rRNA gene-sequence data to estimate microbial diversity based on a cutoff <97% sequence identity for delineation of operational taxonomic units (OTUs). (*A*) Comparison of the composite phyllosphere dataset of this study with published samples covering at least 500 sequences each: farm soil (20), termite gut (19), coastal seawater (17), human gut (18). (*B*) Rarefaction curves of the individual phyllosphere samples and the joint (composite) phyllosphere dataset.

yses as well as in the other 5 samples (Table S2). Further analysis of the clone libraries revealed that between 4% and 10% of the sequences represented unknown genera (see Table S2). Most of them were detected only sporadically, but unknown genera within the family of *Flexibacteraceae* were detected in nearly all samples. Several of the sequences that represented members of known genera were phylogenetically distinct to previously described representatives (type strains) and completely sequenced strains (Fig. S1).

Rarefaction analyses of 16S rRNA gene-sequence data from all 6 samples suggested that the bacterial diversity in the plant phyllosphere samples was lower than in soil, marine systems, or the gut of wood-feeding termites, and similar (Arabidopsis and the Clover 2 sample) or lower (Soybean, Clover 1a and b) than that of the human gut (17–20) (Fig. 3).

Based on cultivation-dependent methods, microbial communities in the phyllosphere have been described to be variable over time, in space, and across different plant species (21, 22). Therefore, DGGE analyses were performed to assess this variation in our field samples. Comparative analysis of the 6 samples showed that similar DGGE patterns were obtained for samples from the same plant species collected at different points in time, suggesting that the bacterial phyllosphere community remained rather stable over time (Fig. S2*a*). This finding was confirmed by the analysis of additional samples taken from the soybean field, which revealed that early colonizers were detectable throughout the whole growing season, while diversity increased during plant succession (Fig. S2*b*). The soybean plant leaves were colonized quite homogenously within the field, as was validated at the time points of harvest of sample material for community proteogenomic analysis (Fig. S2*c*). Taken together, the DGGE analyses showed a temporal and spatial stability of the phyllosphere communities, demonstrating the representativeness of the samples investigated in more detail in the proteome analyses described in the next section.

Comparative Metaproteome Analysis. Proteins from the microbiota of the 6 plant samples were identified after tryptic digestion, using high-accuracy MS. The proteins were processed as described in

Table 1. Identification of abundant proteins in phyllosphere bacteria

	Identifications with RefSeq	Identifications with RefSeq and metagenome	New identifications through metagenome	Gain [%]
Soybean 1	884	934	50	6
Soybean 2	561	1,047	486	87
Clover 1a	556	868	312	56
Clover 1b	442	767	325	74
Clover 2	411	548	137	33
Arabidopsis	505	751	246	49

Gain of protein identifications factored by combining the publicly available database with the generated metagenomic data.

Materials and Methods, and MS/MS spectra were searched against a database consisting of protein sequences obtained from the public RefSeq database with or without the translated metagenomic sequences mentioned above. In total, we identified 2,883 unique proteins with 12,345 peptides, originating from an extensive body of 487,304 spectra (see Table S3 for all identified bacterial proteins and Table S4 for proteins attributed to the respective host plants, soybean- or clover-mosaic viruses, as well as to fungi and oomycetes). The 2,257 bacterial proteins were considered for further interpretation, whereby protein abundance was roughly estimated by spectral counting (23).

The metagenome data significantly increased the number of identified proteins (Table 1), implying the presence of bacteria in our samples that are genetically distinct from those represented among currently sequenced genomes. As expected, the number of identifications increased most strongly for the Soybean 2 sample, from which the metagenome sequences were derived, leading to the identification of 486 additional proteins. Between 6% and 74% of new identifications were obtained for the other 5 samples (see Table 1), a finding that can be ascribed to similarities between bacterial taxa in Soybean 2 and the other samples. An overall consistency of the physiology of the microbiota present on the different plant species is evident at the level of gene expression (i.e., 75% of the proteins identified in the Soybean 2 sample were found in at least 1 of the other samples as well) (Fig. S3).

To assess the significance of similarities and differences in the proteomes and to identify shared and specifically enriched proteins with respect to the 3 different plant species, we examined the identified proteins according to their assignment to Pfam domains (24). This analysis revealed that more than 70% of all identified Pfam domains were present at roughly similar levels on the 3 different plant species (Fig. 4), confirming the overall consistency of the microbiota metaproteomes. Manual inspection of the significantly enriched Pfam-domains (E-value <0.01, P-value <0.0001) revealed that these could most likely be attributed to distinct stresses (as discussed below) or to the presence of distinct bacterial species on the various plant species (see Fig. 4).

Protein Identification in Relation to Bacterial Genera. Most identified proteins were assigned to the 3 bacterial genera *Methylobacterium*, *Sphingomonas*, and *Pseudomonas*, which profited to a different degree from metagenomic information (Table 2, and see Table S3): whereas half of the 20 most abundant proteins of *Methylobacterium* were identifiable through RefSeq and half through the metagenome database, all of the abundant proteins assigned to *Sphingomonas* were identified in various samples based on data we obtained by metagenome sequencing (see Table 2). This suggests that a certain part of the *Methylobacterium* population in the phyllosphere samples is genetically close to the completely sequenced *Methylobacterium* strains currently available in public databases (6 strains), while a major part of the *Sphingomonas* population is different from the sequenced strains (2 strains). These

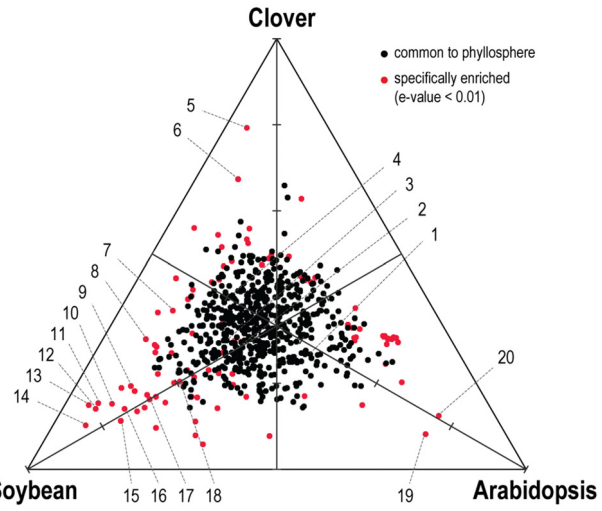


Fig. 4. Conserved and specifically enriched proteome functions (spectral counting of Pfam domains) per host-plant type. Pfam domains drawn close to a vertex are preferentially and specifically found on that respective plant. Selected examples are highlighted and discussed in the text. Examples of common phyllosphere proteome (i.e., not enriched): 1, PF00120, glutamine synthetase catalytic domain; 2, PF02469, fasciclin domain; 3, PF00593, TonB-dependent receptor; 4, PF07715, TonB-dependent receptor plug domain. Specific proteome enrichments: 5, PF00027, cyclic nucleotide-binding domain; 6, PF03328, HpcH/HpaI aldolase/citrate lyase family; 7, PF00210, ferritin-like domain (e.g., bacterioferritins); 8, PF05067, manganese containing catalase; 9, PF06823, protein of unknown function (DUF1236); 10, PF00669, bacterial flagellin N terminus; 11, PF00128, α-amylase, catalytic domain; 12, PF03413, peptidase propeptide and YPEB domain; 13, PF05443, ROS/MUCR transcriptional regulator protein; 14, PF05532, CsbD-like (general stress response); 15, PF00011, Hsp20/alpha crystallin family; 16, PF02566, OsmC-like (e.g., organic hydroperoxide detoxification); 17, PF00700, bacterial flagellin N terminus; 18, PF01584, CheW-like (chemotaxis signaling); 19, PF00532, periplasmic binding and sugar binding domain; 20, PF00502, phycobilisome protein (light harvesting).

conclusions are in agreement with our phylogenetic analysis (see above and Fig. S1). On the other extreme, we observed that all 20 dominant proteins of *Pseudomonas* spp. were identifiable based on RefSeq sequences (see Table 2). This latter observation is also in accordance with our data from 16S rRNA gene-clone library analyses, which showed a very close phylogenetic relationship of the phyllosphere-inhabiting *Pseudomonas* strains to sequenced strains (see Fig. S1*c*). In total, 77 proteins were identified on the basis of metagenome information that did not reveal significant sequence identity to any known or predicted protein (see Table S3). It can, however, not be excluded that some of these are of eukaryotic or viral origin. Notably, 8 of these proteins were found to be expressed in multiple samples among the most abundant proteins (see Table S5). These proteins are of particular interest for further characterization; however, this will most likely require assignment to their respective organisms first.

Plant-Associated Lifestyle

Transport-Related Proteins. Bacterial communities in the phyllosphere are thought to be limited by carbon availability, and it may be expected that access to carbon compounds on leaves is a major determinant of epiphytic colonization (2). There is evidence that small amounts of nutrients, such as simple sugars including glucose, fructose, and sucrose, leach from the interior of the plant (2). We specifically analyzed transport-related functions among the identified proteins to obtain indications for the type of substrates consumed by the phyllosphere microbiota. The most prominent group of transport proteins in our samples consisted of outer-membrane β-barrel proteins (i.e., porins and TonB receptors), which were consistently detected in the analyzed samples from the

Table 2. Most abundant proteins detected in *Methylobacterium*, *Sphingomonas*, and *Pseudomonas*, respectively

Protein (DB)	SY1	SY2	CL1a	CL1b	CL2	ARA
Methylobacterium						
Methanol DH-like XoxF (M)	n.d.	++	+++	+++	+++	++
Fae (M,R)	+++	++	++	++	+++	+
MucR (M)	+	+++	+++	+++	++	+
GroEL (R)	+	++	++	+++	+++	++
Hypothetical protein (R)	++	++	++	+++	++	n.d.
Nucleoside-diP kinase (M)	+	++	++	+++	++	+
Methanol DH MxaF (M,R)	+	+++	++	+++	+	n.d.
Beta-Ig-H3/fasciclin (R)	+++	+++	+	++	+	+
Cold-shock protein (M)	+	++	++	+++	++	+
Beta-Ig-H3/fasciclin (M)	++	+++	+	++	+	+
60 kDa chaperonin (M)	+	+	++	+++	++	n.d.
Phasin (R)	+++	+++	+	+	+	+
Superoxide dismutase (M,R)	+	++	++	++	++	+
Cold-shock protein (M,R)	+	++	++	+	++	+
Chaperonin Cpn10 (R)	++	+	++	++	++	+
Malyl-CoA lyase Mcl (R)	+	+	+	+++	++	+
ClpP (M)	+	+	+++	++	+	+
Surface antigen (M)	n.d.	+	+	++	+++	+
SWIB/MDM2 protein (M)	n.d.	+	++	+	++	n.d.
Invasion associated (M)	n.d.	+	++	++	++	n.d.
Sphingomonas						
OmpA/MotB (M)	+	++	++	+	+	+++
Succinyl-CoA ligase, α (M)	++	+	++	+	+	+++
EF-Tu (M)	+	+	+	++	+	++
OmpA/MotB (M)	n.d.	+	++	++	++	++
EF-Tu (M)	+	+	+	++	+	++
MotA/TolQ/ExbB (M)	+	n.d.	+	+	+	+++
TonB-dependent receptor (M)	n.d.	+	+	++	+	+
GAP dehydrogenase (M)	+	+	+	+	+	+
Histone-like protein (M)	n.d.	+	+	+	+	+
OmpA/MotB (M)	+	++	+	+	n.d.	+
Glutamine synthetase (M)	+	+	+	+	+	++
EF-G (M)	+	+	+	+	+	+
Uncharacterized protein (M)	n.d.	+	+	n.d.	n.d.	++
10 kDa chaperonin (M)	+	+	+	+	n.d.	+
Skp/OmpH (M)	+	+	+	+	n.d.	+
Uncharacterized protein (M)	+	+	n.d.	+	n.d.	+
Membrane protein (M)	+	n.d.	n.d.	+	n.d.	+
TonB-dependent receptor (M)	n.d.	n.d.	n.d.	+	n.d.	+
TonB-dependent receptor (M)	n.d.	n.d.	n.d.	+	+	+
TonB-dependent receptor (M)	+	n.d.	+	+	n.d.	+
Pseudomonas						
OprF (R)	+++	+++	+	n.d.	+	++
Single-stranded binding (R)	+++	++	+	n.d.	+	++
EF-Tu (R)	+++	+	+	n.d.	n.d.	+
Transcript. regulator (R)	+++	+	+	n.d.	n.d.	+
GroEL (R)	+++	+	+	+	+	+
DNA-binding protein (R)	++	+	+	n.d.	+	+
Unknown function DUF883 (R)	++	+	n.d.	n.d.	n.d.	+
Flagellin (R)	+++	+	n.d.	n.d.	n.d.	n.d.
OmpA (R)	++	+	n.d.	n.d.	n.d.	+
F0F1 ATP synthase, β (R)	+	+	+	+	+	+
Succinyl-CoA synth, β (R)	++	n.d.	+	n.d.	n.d.	+
Peptidoglycan lipoprotein (R)	+	+	+	n.d.	n.d.	++
Unknown function DUF883 (R)	++	+	n.d.	n.d.	n.d.	n.d.
Succinyl-CoA synth, α (R)	++	n.d.	+	n.d.	n.d.	n.d.
Chaperone Dank (R)	++	n.d.	+	n.d.	n.d.	n.d.
Glutamine synthetase (R)	++	+	+	n.d.	n.d.	+
Protein P-II (R)	+	n.d.	+	n.d.	n.d.	+
AphC (R)	+	+	+	n.d.	n.d.	+
F0F1 ATP synthase, α (R)	+	+	+	n.d.	+	+
Hsp20 (R)	++	+	n.d.	n.d.	n.d.	n.d.

Proteins were grouped if 90% identical over at least 40% of their length. Taxonomy (at the genus level) was inferred from the protein annotation. Ribosomal proteins are not reported here, but are listed in Table S3. Relative abundances are displayed with +, ++, and +++. DB, database. M (metagenome) and R (Refseq) indicate the database used for identification. n.d., not detected; SY1, Soybean 1; SY2, Soybean 2; CL1a, Clover 1a; CL1b, Clover 1b; CL2, Clover 2; ARA, *Arabidopsis*.

3 different plant hosts. Whereas the former enable passive diffusion of small molecules, the latter allow active transport of substrates greater than ≈600 Da. While we found porins to be abundantly present in various bacterial genera, including *Methylobacterium* and *Pseudomonas*, we observed an over-representation of TonB receptors and the respective plug domains among the proteins assigned to *Sphingomonas* (see Table S3 and Fig. 4). The high number and apparent divergence of the TonB systems is of particular interest, given the rapidly expanding variety of substrates known to be transported by these systems. Beyond the originally identified iron siderophore and vitamin B_{12} transport, the transport of an increasing number of carbohydrates has been reported (25). Our proteome data indicate expression of a gene for a TonB receptor in *Sphingomonas* (see Table S3, identifier Q1NFH3), which is located adjacent to a predicted sucrose hydrolase. Notably, these genes represent orthologs of XCC3358 and XCC3359. XCC3358 was recently described as one of 72 TonB-dependent receptors in the phytopathogen *Xanthomonas campestris* pv. *campestris* (Xcc) transporting sucrose with high affinity, and found to be required for full pathogenicity on *Arabidopsis* (26). Overall, the presence of multiple TonB transporters may account for the large abundance of *Sphingomonas* spp. in terms of abundance on plant leaves by scavenging various substrates present at low amounts, and may reflect a high degree of adaptiveness that can help explain the success of this alphaproteobacterial group.

We also found periplasmic compounds of ABC-transport systems for maltose, glucose, amino acids, and sucrose (see Table S3). Those proteins were more specifically observed to be expressed in *Pseudomonas*, indicating that *Pseudomonas* species could be specialized in mono- and disaccharide utilization and amino acid uptake. Remarkably, only few transporters were assigned to *Methylobacterium* spp.; these consisted mainly of ABC transporters for phosphate and sulfur compounds.

One-Carbon Metabolism. *Methylobacterium* is prominent for its methylotrophic metabolism, which allows it to use methanol, a side product of plant cell-wall metabolism, formed by pectin methyl esterases (27), as its carbon and energy source (28). The presence of this metabolic ability was suggested by numerous highly abundant proteins (see Table 2), including the large subunit of the periplasmic pyrrolo quinoline quinone-containing methanol dehydrogenase (MxaF) and a complete set of proteins of the tetrahydromethanopterin-dependent pathway (29). Moreover, proteins involved in the assimilation of methanol-derived methylene tetrahydrofolate and carbon dioxide via the serine pathway were detected, such as serine-glyoxylate aminotransferase, hydroxypyruvate reductase, and malyl-CoA lyase (30). These proteins are essential for methylotrophic growth and the encoding genes are located in a large genomic region (30), which is displayed in Fig. S4 together with identified peptides.

This genomic methylotrophy region also contains a gene for a methanol dehydrogenase-like protein (XoxF), which exhibits a sequence identity of 50% to MxaF. Under laboratory culture conditions, we were able to detect only very little of this protein in *Methylobacterium extorquens* cells and Bosch et al. (31) determined a 100-fold lower expression of *xoxF* compared to *mxaF* based on spectra counting of peptides. So far, no phenotype was observed for a *xoxF* mutant in *M. extorquens* AM1 (32) (for occurrence of *xoxF* and assumed functions in other bacteria see ref. 33). In contrast, upon plant colonization *xoxF* is highly expressed in *Methylobacterium* (see Table 2). For an approximation of expression levels, we integrated and correlated metagenomic and metaproteomic information using a 2-way fragment-recruitment approach, which revealed that the expression of *xoxF* was roughly in the same range as that for *mxaF* (Fig. S5). In the Arabidopsis sample, XoxF was even detected exclusively; that is, no MxaF was detectable. The high expression level of *xoxF* in *Methylobacterium* under environmental conditions suggests an important physiological role of XoxF during

plant colonization. Further analyses of this protein, in particular with regard to substrate specificity and affinity, will be of great interest.

Overall, the detection of proteins known to be involved in methylotrophy and their assignment to *Methylobacterium* spp. suggests that facultative Methylobacteria are the dominating methylotrophs on plants, and that the large success of these bacteria in the phyllosphere can likely be attributed to specialization in carbon source utilization.

Nitrogen Metabolism. Bacteria can use various nitrogen sources, including ammonia, nitrate, dinitrogen, and a variety of amino acids and other nitrogenous organic compounds. The amino acid transporters mentioned above suggest that plant-derived nitrogen compounds are available for the bacteria. In addition, ammonia may be used as a nitrogen source, as suggested by the prominent presence of glutamine synthetase (see Fig. 4) in various bacteria, including *Sphingomonas*, *Methylobacterium*, and *Pseudomonas*. Indications for a dinitrogen fixation ability among the identified proteins of the phyllosphere microbiota inhabiting the studied plants were not found.

Stress Resistance. The phyllosphere is known as a hostile environment for the residing microorganisms (2, 9). In addition to the oligotrophic character of this habitat, physical parameters contribute to stressful conditions, such as UV radiation, temperature shifts, and the presence of reactive oxygen species. Adaptation to stressful conditions was reflected by the detection of various proteins, assigned to diverse bacterial genera and detected in all analyzed samples. Among these proteins were superoxide dismutase, catalase, DNA protection proteins, chaperones, and proteins involved in the formation of the osmoprotectant trehalose. Recently, evidence was presented that general stress response is an essential mechanism for plant colonization by *Methylobacterium* (9, 34). The regulatory system of general stress response in *Methylobacterium*, and presumably in other Alphaproteobacteria, consists of the 2-component response regulator PhyR that triggers upon activation regulation of stress-related protein functions via sigma factors of the EcfG family (35). PhyR and EcfG, respectively, were found among the detected proteins within this study (see Table S3) from members of the alphaproteobacterial genera *Methylobacterium*, *Sphingomonas*, and *Aurantimonas*, thus further emphasizing the importance of these regulatory proteins.

For *Pseudomonas*, besides the stress-response proteins, such as alkyl hydroperoxide reductase, DNA protection proteins, catalase, and the periplasmic serine protease MucD, a number of regulators were identified that are known to be related to stress response in this Gammaproteobacterium. These regulators were the oxidative stress-response regulator OxyR, and regulators such as AlgR, AlgR3, and AlgU (AlgT) (see Table S3). The latter belongs to the ECF-family of sigma factors and regulates *algD* expression. The AlgD protein, which was also detected in this study (see Table S3), is involved in biosynthesis of the exopolysaccharide alginate, which has been demonstrated to be of importance for increased epiphytic fitness, virulence, and resistance to desiccation and toxic molecules (36).

An over-representation of stress-related proteins was found in the soybean samples (see Fig. 4). This might reflect a consequence of a plant-defense response, which in turn was possibly triggered by the presence of flagellin (37) of *Pseudomonas* spp. (see below). Strains with very close relationship to the pathogen *P. syringae* pv. *glycinae* (100% sequence identity on 16S rRNA gene level) were detected on the soybean plants.

Motility. We observed a significant over-representation of flagellin in *Pseudomonas* relative to other bacteria (see Table S3, Table 2, Fig. 4 and Fig. S4). It is conceivable that *Pseudomonas* spp. rather than *Methylobacterium* spp. and *Sphingomonas* spp. have adapted a lifestyle that is predestinated to actively search for nutrients. Motility is well established as an important epiphytic fitness factor of plant colonizing *Pseudomonas* (38) and was shown to be regulated by quorum sensing (39). Apparently, *Pseudomonas* spp. are not part of the common and consistent microbiota on plants, but rather transient inhabitants probably subjected to more frequent changes in abundance (see Table S2) (see also refs. 21 and 40).

Conspicuous Proteins. Finally, we searched the metaproteomic dataset for the presence of proteins of unknown or poorly characterized function that were consistently present throughout our samples and among different bacterial species, as they may be indicative for a common trait shared by bacteria adapted to the phyllosphere. Among these proteins, "beta-Ig-H3/fasciclin" was prominent (see Table 2 and Fig. 4). Proteins of this family were detected based on genome sequence information from *Methylobacterium* (see Table 2 and Fig. S4), *Rhodopseudomonas*, *Novosphingobium*, and *Stenotrophomonas* among the most abundant proteins identified in this study (see Table S5), and from a number of other bacterial genera when considering all identified proteins (see Table S3). Homologues of this fasciclin domain protein are found in vertebrates and invertebrates and are thought to mediate cell adhesion (41). Notably, fasciclin homologues were described to be symbiotically relevant in 3 separate cases (*Nostoc*–lichens, *Rhizobium*–legume, and algae–cnidaria) (42, 43). Consequently, the fasciclin protein is a prime candidate for further investigation with regard to its importance for bacteria during the phyllospheric lifestyle and its putative role in cell-cell adhesion. Another example of a consistently detected protein in several bacterial species is given in Fig. S4 (TypA/BipA).

Conclusions

To our knowledge, this study is innovative in representing a large-scale combinatorial metagenome and metaproteome analysis from a common pool of cells. This approach allowed us to overcome limitations in protein identification that are otherwise encountered because of the absence of closely related reference genomes in publicly available databases. It also demonstrated that metagenome data, retrieved from relatively short sequence reads and with low degree of assembly, are of sufficient quality to allow protein identification of bacteria not sequenced so far. The identification of abundant proteins in the phyllosphere microbiota allowed us to detect key enzymatic functions with activities that can be expected to be relevant for global carbon and nitrogen cycles. This holds especially for the conversion of methanol, a major volatile organic compound emitted by plants (100 Tg formed per year) (27), and the assimilation of ammonia via glutamine synthetase. The latter is of relevance considering the high amount of ammonia input from agricultural sources and from industrial exhaust, as discussed in relation to the phyllosphere (44).

The identity of bacteria present in the phyllosphere in combination with the protein survey described here offers insights into strategies for phyllospheric lifestyles of bacteria on plant hosts. Our analysis revealed consistency with respect to the bacterial community composition and, in particular, the high abundance of *Sphingomonas* spp. and *Methylobacterium* spp. on the analyzed plants. Known proteins expressed in *Methylobacterium* are related, to a large extent, to one-carbon and central metabolism, as well as to stress response, whereas for *Sphingomonas* spp., the conspicuous expression of TonB-dependent receptors suggests a particularly large substrate spectrum. These adaptations contribute to the success and coexistence of these taxa in the phyllosphere. Apart from these consistently observed 2 alphaproteobacterial genera, we detected the presence of flagellated *Pseudomonas* on soybean plants and with it a number of proteins of known and unknown functions.

The survey of proteins present in situ provides a basis for targeted studies of proteins relevant in relation to the plant

environment. Strikingly, the consistent and abundant presence of some proteins of uncharacterized function in a number of different bacterial genera, of which fasciclin is one example, suggest key functions for adaptation to the phyllosphere that need to be investigated in more detail. The identity of abundant and ubiquitous commensal phyllosphere bacteria in combination with a better understanding of their physiology in this habitat will help to reveal the role of these bacteria in global carbon and nitrogen cycles, and serve as a basis to exploit them in the future with respect to a potential plant probiotic power.

Materials and Methods

Sampling of Phyllosphere Bacteria and Extraction of DNA and Protein. Bacterial cells were washed from the leaf material applying a previously published protocol (9) with slight modifications (see *SI Text*), including a centrifugation step in the presence of Percoll to deplete eukaryotic cells and dirt particles. DNA and protein extraction was performed using the AllPrep DNA/RNA/Protein Mini Kit (Qiagen). Frozen cell pellets were resuspended in 1,300 to 1,400 μl of kit-supplied RLT buffer, 1 g of 0.1-mm zirconium-silica beads was added, and cell lysis was performed in a tissue lyser (Retsch GmbH) for 3 min at maximum shaking frequency (30 s^{-1}). Cell debris and beads were pelleted for 1 min at 20,000 $\times$ *g*. The supernatant was distributed onto 2 kit-supplied columns for further extraction of the DNA and proteins according to the instructions in the kit manual.

DNA Metagenome Sequencing and Analysis. Sequencing was performed on the Genome Sequencer FLX system. All DNA sequences were assembled with the GS De Novo Assembler provided with the FLX system (Roche Applied Science and 454 Life Sciences) using default parameters for protein identification. ORFs were predicted and data annotated as outlined in the *SI Text*. Taxonomic community composition estimates based on metagenomic sequences were derived by running the software MLTreeMap on the Soybean 2 metagenomic data (13).

Microbial Community 16S rRNA Gene-Based Analysis. The bacterial and archaeal community composition of the 6 phyllosphere samples was characterized by 16S rRNA gene-clone library construction, followed by comparative sequence analysis as outlined in detail in the *SI Text*. Rarefaction curves were calculated using the Dotur software package (45).

Protein Identification and Analysis. Proteins were separated by 1-dimensional SDS/PAGE and analyzed after tryptic digestion by reversed-phase high-performance liquid-chromatography coupled to electrospray-ionization tandem mass-spectrometry. Data files obtained from high-accuracy mass spectrometers were converted to peak lists and were analyzed with 2 search algorithms and validated with Scaffold (Proteome Software Inc.). MS/MS spectra were searched against 2 different databases: one database consisting of protein sequences obtained from RefSeq (ftp://ftp.ncbi.nih.gov/refseq) and a second database built from RefSeq data plus the translated metagenomic data (see Dataset S1). For protein identification, at least 2 peptide matches were required (each having a minimum peptide identification probability of 95%; minimum required protein identification probability was 99%). The false discovery rate, as estimated by searches against a decoy database, was below 1%. Data processing and visualization were performed using custom scripts in Perl, Python, and R. Full information about all of the methods and associated references used for the analyses reported here is available in the *SI Text*.

ACKNOWLEDGMENTS. We thank Carlos Alonso Blanco (Spanish National Center for Biotechnology), Thomas Hebeisen, Daniel Suter, and Christine Herzog (Forschungsanstalt Agroscope Reckenholz-Tänikon, Switzerland) for support with the plant sampling, Marzanna Künzli (Functional Genomics Center Zurich) for support with genome sequencing, Simon Barkow-Oesterreicher (Functional Genomics Center Zurich) for support with database handling, and Roger Wepf (Electron Microscopy Center of the Eidgenössische Technische Hochschule Zurich) for support with electron microscopy imaging. We thank the Vital-IT group of the Swiss Institute of Bioinformatics for providing computational resources. The work was supported by Eidgenössische Technische Hochschule Zurich and by the University of Zurich through its Research Priority Program ''Systems Biology and Functional Genomics''.

1. Bailey MJ (2006) *Microbial Ecology of Aerial Plant Surfaces* (CABI Publishing, Wallingford).
2. Lindow SE, Brandl MT (2003) Microbiology of the phyllosphere. *Appl Environ Microbiol* 69:1875–1883.
3. Lambais MR, Crowley DE, Cury JC, Bull RC, Rodrigues RR (2006) Bacterial diversity in tree canopies of the Atlantic forest. *Science* 312:1917.
4. Redford AJ, Fierer N (2009) Bacterial succession on the leaf surface: A novel system for studying successional dynamics. *Microb Ecol* 58:189–198.
5. Yang CH, Crowley DE, Borneman J, Keen NT (2001) Microbial phyllosphere populations are more complex than previously realized. *Proc Natl Acad Sci USA* 98:3889–3894.
6. Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. *Annu Rev Microbiol* 57:369–394.
7. Leveau JHL (2006) Microbial communities in the phyllosphere. In *Biology of the Plant Cuticule*, eds Riederer M, Müller C (Blackwell, Oxford), pp 334–367.
8. Boch J, et al. (2002) Identification of *Pseudomonas syringae* pv. *tomato* genes induced during infection of *Arabidopsis thaliana*. *Mol Microbiol* 44:73–88.
9. Gourion B, Rossignol M, Vorholt JA (2006) A proteomic study of *Methylobacterium extorquens* reveals a response regulator essential for epiphytic growth. *Proc Natl Acad Sci USA* 103:13186–13191.
10. Marco ML, Legac J, Lindow SE (2005) *Pseudomonas syringae* genes induced during colonization of leaf surfaces. *Environ Microbiol* 7:1379–1391.
11. Yang S, et al. (2004) Genome-wide identification of plant-upregulated genes of *Erwinia chrysanthemi* 3937 using a GFP-based IVET leaf array. *Mol Plant Microbe Interact* 17:999–1008.
12. VerBerkmoes NC, Denef VJ, Hettich RL, Banfield JF (2009) Systems biology: Functional analysis of natural microbial consortia using community proteomics. *Nat Rev Microbiol* 7:196–205.
13. von Mering C, et al. (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* 315:1126–1130.
14. Corpe WA, Rheem S (1989) Ecology of the methylotrophic bacteria on living leaf surfaces. *FEMS Microbiol Ecol* 62:243–250.
15. Kim H, et al. (1998) High population of *Sphingomonas* species on plant surface. *J Appl Microbiol* 85:731–736.
16. Knief C, Frances L, Cantet F, Vorholt JA (2008) Cultivation-independent characterization of *Methylobacterium* populations in the plant phyllosphere by automated ribosomal intergenic spacer analysis. *Appl Environ Microbiol* 74:2218–2228.
17. Acinas SG, et al. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* 430:551–554.
18. Eckburg PB, et al. (2005) Diversity of the human intestinal microbial flora. *Science* 308:1635–1638.
19. Hongoh Y, et al. (2005) Intra- and interspecific comparisons of bacterial diversity and community structure support coevolution of gut microbiota and termite host. *Appl Environ Microbiol* 71:6590–6599.
20. Tringe SG, et al. (2005) Comparative metagenomics of microbial communities. *Science* 308:554–557.
21. Ellis RJ, Thompson IP, Bailey MJ (1999) Temporal fluctuations in the pseudomonad population associated with sugar beet leaves. *FEMS Microbiol Ecol* 28:345–356.
22. Kinkel LL (1997) Microbial population dynamics on leaves. *Annu Rev Phytopathol* 35:327–347.
23. Liu H, Sadygov RG, Yates JR, 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* 76:4193–4201.
24. Finn RD, et al. (2008) The Pfam protein families database. *Nucleic Acids Res* 36:D281–D288.
25. Schauer K, Rodionov DA, de Reuse H (2008) New substrates for TonB-dependent transport: do we only see the 'tip of the iceberg'? *Trends Biochem Sci* 33:330–338.
26. Blanvillain S, et al. (2007) Plant carbohydrate scavenging through TonB-dependent receptors: a feature shared by phytopathogenic and aquatic bacteria. *PLoS ONE* 2:e224.
27. Galbally IE, Kirstine W (2002) The production of methanol by flowering plants and the global cycle of methanol. *J Atmosph Chem* 43:195–229.
28. Sy A, Timmers AC, Knief C, Vorholt JA (2005) Methylotrophic metabolism is advantageous for *Methylobacterium extorquens* during colonization of *Medicago truncatula* under competitive conditions. *Appl Environ Microbiol* 71:7245–7252.
29. Vorholt JA (2002) Cofactor-dependent pathways of formaldehyde oxidation in methylotrophic bacteria. *Arch Microbiol* 178:239–249.
30. Chistoserdova L, Chen SW, Lapidus A, Lidstrom ME (2003) Methylotrophy in *Methylobacterium extorquens* AM1 from a genomic point of view. *J Bacteriol* 185:2980–2987.
31. Bosch G, et al. (2008) Comprehensive proteomics of *Methylobacterium extorquens* AM1 metabolism under single carbon and nonmethylotrophic conditions. *Proteomics* 8:3494–3505.
32. Chistoserdova L, Lidstrom ME (1997) Molecular and mutational analysis of a DNA region separating two methylotrophy gene clusters in *Methylobacterium extorquens* AM1. *Microbiology* 143:1729–1736.
33. Chistoserdova L, Kalyuzhnaya MG, Lidstrom ME (2009) The expanding world of methylotrophic metabolism. *Annu Rev Microbiol* 63:477–499.
34. Gourion B, Francez-Charlot A, Vorholt JA (2008) PhyR is involved in the general stress response of *Methylobacterium extorquens* AM1. *J Bacteriol* 190:1027–1035.
35. Francez-Charlot A, et al. (2009) Sigma factor mimicry involved in regulation of general stress response. *Proc Natl Acad Sci USA* 106:3467–3472.
36. Yu J, Penaloza-Vazquez A, Chakrabarty AM, Bender CL (1999) Involvement of the exopolysaccharide alginate in the virulence and epiphytic fitness of *Pseudomonas syringae* pv. *syringae*. *Mol Microbiol* 33:712–720.
37. Boller T, He SY (2009) Innate immunity in plants: an arms race between pattern recognition receptors in plants and effectors in microbial pathogens. *Science* 324:742–744.
38. Haefele DM, Lindow SE (1987) Flagellar motility confers epiphytic fitness advantages upon *Pseudomonas syringae*. *Appl Environ Microbiol* 53:2528–2533.
39. Quinones B, Dulla G, Lindow SE (2005) Quorum sensing regulates exopolysaccharide production, motility, and virulence in *Pseudomonas syringae*. *Mol Plant Microbe Interact* 18:682–693.
40. Hirano SS, Upper CD (2000) Bacteria in the leaf ecosystem with emphasis on *Pseudomonas syringae*—a pathogen, ice nucleus, and epiphyte. *Microbiol Mol Biol Rev* 64:624–653.
41. Carr MD, et al. (2003) Solution structure of the *Mycobacterium tuberculosis* complex protein MPB70: from tuberculosis pathogenesis to inherited human corneal disease. *J Biol Chem* 278:43736–43743.
42. Oke V, Long SR (1999) Bacterial genes induced within the nodule during the *Rhizobium*-legume symbiosis. *Mol Microbiol* 32:837–849.
43. Paulsrud P, Lindblad P (2002) Fasciclin domain proteins are present in *Nostoc* symbionts of lichens. *Appl Environ Microbiol* 68:2036–2039.
44. Papen H, Gessler A, Zumbusch E, Rennenberg H (2002) Chemolithoautotrophic nitrifiers in the phyllosphere of a spruce ecosystem receiving high atmospheric nitrogen input. *Curr Microbiol* 44:56–60.
45. Schloss PD, Handelsman J (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* 71:1501–1506.

Supporting Information

Delmotte et al. 10.1073/pnas.0905240106

SI Materials and Methods

Harvest of Prokaryotic Phyllosphere Cells. Plant leaf material [i.e., rosettes of thale cress (*Arabidopsis thaliana*), fully developed leaves of soybean (*Glycine max*), or fully developed trifoliates of clover (*Trifolium repens*)] was placed in 50-mL tubes and the tubes were filled up to 30 ml with sterile, precooled TE-buffer (10 mM Tris, 1 mM EDTA, pH 7.5), supplemented with 0.3 g mL^{-1} Pefabloc SC (Roche Diagnostics) and 0.2% Silwet L-77 (GE Bayer Silicones). Cells were washed from the leaf material by 3 min of alternate shaking, vortexing, and sonication. The cell suspension was separated from the leaf material by filtration through a nylon mesh (pore size, 200 μm; Spectrum Europe BV). Six milliliters of 80% Percoll (Sigma-Aldrich) was pipetted below the cell suspension, and the 50-mL tubes were centrifuged for 5 min at 800 $\times$ *g*. The bacterial cell suspension above the Percoll layer was transferred into a fresh 50-mL tube and cells were pelleted at 3,150 $\times$ *g* for 15 min. Cell pellets from multiple tubes were pooled into 1.5-ml reaction tubes and washed twice with TE-buffer plus Pefabloc SC. Cell pellets were immediately frozen at –20 °C.

Construction and Analysis of 16S rRNA Gene-Clone Libraries. Bacterial 16S rRNA genes were amplified in triplicate PCR assays (volume of 33 μl each, prepared from a master mix of 100 μl). Each 100-μl assay contained 10 μl of supplied RedAccu LA Taq Polymerase PCR buffer containing 2.5 mM of Mg^{2+} (Sigma), 1.25 mM of each deoxynucleoside triphosphate (dNTP) (Fermentas), 0.5 μM of each primer (Microsynth), 0.25 μg μL^{-1} of BSA (Roche Diagnostics), 0.05 U μL^{-1} of Red Accu LA Taq polymerase (Sigma), and 5 μl of template DNA. Primers 9f and 1492r were used for PCR amplification of bacteria (1). The PCR program consisted of initial denaturation at 94 °C for 4 min, followed by 25 cycles of denaturation at 94 °C for 45 s, annealing at 48 °C for 1 min, and elongation at 72 °C for 2 min, and then a final elongation at 72 °C for 7 min. PCR products were purified with a NucleoSpin Extract II purification kit (Machery-Nagel), and A-overlaps were replenished in an assay containing 5-μl purified PCR product, 0.6 μl of supplied Master Taq Polymerase buffer (Eppendorf), 0.3 μl of each dNTP, and 0.3 μl of Master Taq Polymerase (Eppendorf) by incubation at 72 °C for 10 min. For the detection of Archaea, a specific primer system was applied (20f + 958r) (2). PCR was performed in assays as described above using the thermal profile as described (2) with 35 reaction cycles. PCR products could be obtained in the Soybean 1, Soybean 2, Clover 2, and *A. thaliana* samples.

After cloning and sequencing with primers 9f and 1492r, the nearly full-length 16S rRNA gene sequences were obtained after assembly and aligned using the SINA webaligner of the SILVA ribosomal database project (3). Sequences were double-checked for chimeras using the Mallard program and the chimera detection program of the ribosomal database project RDP, release 8. Sequences that showed anomalies with only 1 of the 2 programs were manually checked with Pintail. Moreover, the aligned sequences were visually inspected for anomalies.

Phylogenetic trees were calculated using the maximum-likelihood algorithm PhyML, implemented in the ARB software package. Type strains for the trees shown in Fig. S1 *a–d* were selected according to "The All-Species Living Tree project," release 93 (4).

Denaturing Gradient Gel Electrophoresis. DGGE was performed as previously described (5). Briefly, primers 533f and 907r-GC were applied to PCR-amplify a fragment of the 16S rRNA gene with 35 cycles. PCR products were quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen), and equal amounts of DNA were loaded onto each gel. Acrylamide gels (6.5%) were prepared with a denaturing gradient from 35 to 65%, and gels were run at 60 °C and 70 V for 16 h. Excised bands were reamplified with 25 PCR cycles and the correct migration behavior was checked on a DGGE before sequencing. The community composition of the 6 samples used for metaproteomic analysis was additionally analyzed by using a second primer system, 357f-GC and 907r (6), which revealed a comparable clustering of samples (i.e., samples from the same plant species clustered together). DGGE patterns were compared with the GelCompar II software (Applied Maths). Cluster analysis was performed using the unweighted-pair group method using arithmetic averages algorithm based on Pearson correlation coefficients.

DNA Metagenome Sequence Analysis. Pyrosequencing was performed by GATC and at the Functional Genomics Center Zurich using an aliquot from the DNA extract of the Soybean 2 sample. Five micrograms of DNA was provided for each analysis. DNA quantity and purity (based on the ratio of absorbance at 260 and 280 nm and was 1.7 for our sample) was determined using the NanoDrop (Thermo Fisher Scientific) and the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen). Data assembly using the GS De Novo Assembler resulted in 140,550 contigs with a mean sequence length of 276 bp, or 40-bp longer than the mean length of a single read. The largest contig had a length of 12,888 bp. After assembly, different read statuses were attributed to each read by the assembler software: assembled, partially assembled (only part of the read included), singleton (no overlap with any other read), repeat (identified as a repeat region or exactly duplicated sequence; known artifact of the pyrosequencing technique), or outlier (problematic read; for example, chimera sequences). To build the metagenome database (for proteomic data annotation), singleton reads were included in the contigs file in order not to lose any information after assembly. The annotation of contigs and singleton reads was performed as follows: ORF prediction, with translation of regions between stop codons in the 6 reading frames, was done using the program *getorf* (EMBOSS package). ORFs with a minimum size of 10 aa were reported. Similarity searches for all predicted ORFs were performed using the program BLASTp with an expected (E) value cutoff of 0.0001 (and the following parameters: "-M BLOSUM62 -G 11 -E 1 -F T") against the database UniRef90. A hit was considered significant with a bitscore larger than or equal to 60. Pfam domains (7) were reported for ORFs that significantly matched UniRef90 using the mapping file *protein2ipr.dat.gz* available on the Interpro ftp Web site. A domain was reported if containing a minimum overlap of 20 aa with the contig/read. A total of 319,651 ORFs matched those criteria. All nonannotated ORFs (5,647,279) were kept in the metagenome database (total of 5,966,930 entries) for further analysis in case of identification by MS.

Preparation of Proteins for MS. The extracted protein fraction of each sample, obtained as indicated above, was processed further using the Allprep DNA/RNA/Protein kit (Qiagen). Proteins were precipitated and then dissolved in a Laemmli-related kit-supplied sample buffer. If needed, proteins were frozen and stored at −20 °C; otherwise, the proteins were diluted up to 45

μl in loading buffer and denaturated for 4 min at 95 °C. Loading buffer was prepared by mixing 125 μl of 0.5 M Tris-HCl, pH 6.8, 250 μl of glycerol, 200 μl of 10% SDS, 50 μl of 2-β-mercaptoethanol, and 1 crystal of bromophenol blue and then bringing the solution to a final volume of 2 ml with water. After cooling and centrifugation at 20,238 $\times$ *g* for 5 min, the protein sample was loaded for separation on the top of a Tris-HCl polyacrylamide gel (4–15% linear gradient, 8.6 $\times$ 6.8 cm, or 10.5–14% linear gradient, 13.3 $\times$ 8.7 cm) obtained from Bio-Rad Laboratories AG. Electrolysis buffer consisted of 25 mM Tris-HCl, pH 8.3, 192 mM glycine, and 0.1% SDS. Staining was performed for 40 min with 40% methanol, 10% acetic acid, and 0.25% Coomassie blue. Destaining was achieved overnight with 10% methanol and 10% acetic acid. For each sample, the corresponding gel lane was cut into 16 to 21 pieces. Gel pieces were destained 3 times with 50% acetonitrile and dried for 10 min under vacuum (Model SPD121P SpeedVac, Thermo Fisher Scientific). Then, proteins were reduced for 45 min at 56 °C with Tris(2-carboxyethyl)phosphine hydrochloride (2 mM in 25 mM ammonium hydrogen carbonate, pH 8.0) and carbamidomethylated for 60 min at room temperature in the dark with iodoacetamide (25 mM in 25 mM ammonium hydrogen carbonate, pH 8.0). Gel plugs were washed 3 times with 50% acetonitrile and dried for 15 min under vacuum. Finally, proteins were digested with trypsin (Promega) for 16 h at 37 °C (50 ng/gel plug) in 25-mM ammonium hydrogen carbonate, pH 8.0. Digestion was quenched with trifluoroacetic acid, digests were transferred to new vessels and solvents were evaporated. After resolubilisation in 30 μl of 3% acetonitrile and 0.1% trifluoroacetic acid, peptides were cleaned up with a C18 ZipTip supplied by Millipore Corporation.

MS Analysis. The samples were analyzed on a hybrid LTQ-Orbitrap XL mass spectrometer (Thermo Fisher Scientific) interfaced with a nanoelectrospray source. Peptides were separated by reversed-phase high-performance liquid-chromatography on an in-house packed column with 2 μm UltraHT Pro C18 packing material from YMC Co. Column dimensions were 80 $\times$ 0.75 mm inside diameter. Eluents were (A) 1% acetonitrile, 0.2% formic acid, and (B) 80% acetonitrile, 0.2% formic acid. Separation was performed by linear gradients of 3 to 10% (B) in 5 min, 10 to 40% (B) in 50 min, 40 to 97% (B) in 5 min, followed by isocratic conditions at 97% (B) for 5 min. Solvent delivery of 200 nL min^{-1} was achieved by a binary gradient pump (Model nanoLC 1D Plus, Eksigent). Peptides were loaded from a cooled (4 °C) auto sampler (Model Endurance, Spark Holland). Connection of the reversed-phase column with the ESI source was achieved by stretching the fused silica capillary at the outgoing extremity of the column.

MS detection was performed with the LTQ-Orbitrap XL mass spectrometer operating in data-dependent mode. The 4 most abundant doubly or triply charged ions from the high-accuracy survey scan with a minimum ion count of 500 were automatically taken for further MS/MS analysis at the linear ion trap. Precursor masses already taken for MS/MS were excluded for further selection for 60 s. All mass spectra were recorded in positive ion mode with an electrospray source voltage between 1.5 kV and 1.90 kV. Precursor mass spectra were acquired at the Orbitrap mass analyzer with a scan range from *m/z* 300.0 to 1,600.0 using real-time internal calibration on polydimethylcyclosiloxane background ions *m/z* 445.120025 and 429.088735, as previously described (8). Resolution was set to 60,000 at *m/z 400*. For some remeasurements, a hybrid LTQ-FTICR mass spectrometer (Model LTQ-FT Ultra, Thermo Fisher Scientific) was used. Chromatographic separation, ionization, and data acquisition were performed as described for the LTQ-Orbitrap XL mass spectrometer.

Protein Identification and Determination of False-Discovery Rate. Mass spectra processing was performed with Xcalibur 2.0.7 (Thermo Fisher Scientific). Peak list generation for database searches was performed with Mascot Distiller 2.1.1.0 (Matrix Science). Database searches were performed against 3 different databases. The first database (DB1), containing 5,195,116 protein sequences, consisted of RefSeq Release 28 and was downloaded from the NCBI ftp Web site (ftp://ftp.ncbi.nih.gov/refseq/release). The second database (DB2) was a concatenation of DB1 with 5,966,930 sequences issued from the metagenomics part of the project and had a total of 11,162,046 entries. The third database (DB3) had 15,285 entries and consisted of all of the protein sequences from 3 reference complete genomes, *Methylobacterium extorquens* PA1, *Sphingomonas wittichii* RW1, and *Pseudomonas syringae* pv. *phaseolicola* 1448A, downloaded from the NCBI ftp Web site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). The selection of reference genomes was based on number of identified proteins for different species within the genus and was thus based on results shown in Table S3.

For database searches, a first computation was performed with Mascot 2.2 (Matrix Science) based on the MOWSE algorithm (9). The following search parameters were applied: taxonomy, all entries; fixed modification, cysteine carbamidomethylation; variable modifications, methionine oxidation; enzyme, trypsin; maximum number of missed cleavages, 1; peptide tolerance, $\pm$ 5 ppm; MS/MS tolerance, $\pm$ 0.5 Da. We were able to set a low peptide tolerance (5 ppm) because of the high accuracy of the Orbitrap mass spectrometer and the use of internal lock-mass calibration with the polydimethylcyclosiloxane background ions. By acquiring the data with high accuracy we were able to obtain peptide matches with high MOWSE scores (high quality), which are above identity cutoffs computed by Mascot (typically 40 with huge databases). A second database search was performed by using the X!Tandem database searching program (10). Results from both algorithms were validated with Scaffold 2.1 (Proteome Software Inc.). Peptide identifications were accepted if they could be established at greater than 95% probability as specified by the Peptide Prophet algorithm (11). Probability [in the sense of the Protein Prophet algorithm (12)] greater than 99% was required to validate protein identifications. One-hit wonders were removed (only proteins identified with at least 2 peptides were considered) and proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principle of parsimony.

To check the quality of our validation process, we prepared a synthetic mixture of 10 bacteria occurring in environmental samples. Both gram-positive and gram-negative species were represented and protein concentrations varied over 2 orders of magnitude between the different species. The protein mixture was processed as described for the real samples. Mass spectra were searched against DB1. Less than 1% of the hits we obtained were false-positive.

We also computed the false-discovery rate by testing experimental mass lists against a composite version of database DB2, created by concatenating the target protein sequences with reversed sequences (total of 22,324,092 sequences, target-decoy searches) as described by Elias and Gygi (13). Because we searched the mass lists against a database containing forward and reverse sequences, the number of identified reverse hits was multiplied by 2 and divided by the total number of identifications. We computed a false-discovery rate lower than 1%.

Contrary to classical proteomics, for which protein assignment to a given organism is obvious, protein assignment to a given taxon in community proteomics may remain uncertain (see also ref. 14).

Spectral Counting. Given the redundancy and diversity of identified proteins using the database DB2, we performed a clustering of the corresponding sequences to facilitate the interpretation of the results and to be able to roughly estimate protein expression by spectral counting (15). A single linkage clustering based on sequence identity was performed using the program BLASTCLUST and the following parameters "-p T -e F -L .4 -b T -S 90." Sequences aligning at least 40% of their length and with an identity superior or equal to 90% were clustered together. For a given sample, the cluster spectral count (the sum of spectral counts for all of the proteins in a given cluster) was normalized according to the total number of spectra acquired for this sample. Because longer proteins have a greater chance to be detected via MS, we also normalized cluster spectral counts by the longest protein length present in a cluster. Finally, we report the normalized spectral counts as +++, ++, and + for values ≥ 1.7, ≥ 0.9 and <1.7, and <0.9, respectively.

To better characterize clusters, biologically and functionally, we annotated them using the Gene Ontology database (http://www.geneontology.org/) using precomputed annotation available for Uniprot proteins in the GOA database (http://www.ebi.ac.uk/GOA/) and the online Protein Identifier Cross-Reference Service (http://www.ebi.ac.uk/Tools/picr/).

Differential Proteome Composition. The similarity between plant-sample proteomes was analyzed based on expressed Pfam protein domains. Spectral counting was performed to semiquantitatively estimate protein abundance. For each known protein domain (Pfam), these abundances were then aggregated based on the protein/domain mappings (in this case, we used the whole length of the identified protein; that is, even for cases where a given domain was not itself covered by peptides, it clocked counts based on the annotated occurrence of that domain in the protein). To investigate which protein domains were consistently expressed or plant-specifically enriched, we pooled the samples according to the plant species. A triangular representation was used to visualize the specific enrichments of domains detected on each plant species. Each protein domain is represented by 1 dot within the triangle, whereby the position of the dot signifies the relative enrichment of the domain in one or several of the samples. Domains that are equally frequent on all 3 plants appear in the middle of the triangle. Domains that appear in 1 of the corners of the triangle are found primarily on 1 of the plants, and domains that appear along 1 of the edges of the triangle are found primarily in 2 of the 3 sample pools, but are largely absent from the third. For each protein domain, the relative counts for the 3 habitats were normalized to add up to 1 (after addition of pseudocounts to select against rare domains). This permitted the display of 3-dimensional data in 2 dimensions (using 3 axes at 120° angles). Statistical significance assessment was performed using a Monte-Carlo method (comparison to randomized data). For more details on the method, see Tringe et al. (16).

Two-Way Fragment Recruitment. The fragment recruitment analysis was developed using a custom Python script to integrate the data and generate fragment recruitment plots. The DNA short reads recruitment was performed using the program BLAST (17) and the following parameters "a 3 -F 'L;m;' -e 0.0001 -G 5 -E 2 -r 2." All 454 reads were searched for similarity against a database (DB3) containing the 3 reference genomes (*Methylobacterium extorquens* PA1, *Sphingomonas wittichii* RW1, and *Pseudomonas syringae phaseolicola* 1448A) and their respective plasmid sequences downloaded from the RefSeq database. Best hits on a given genome were defined by the best bitscore and a bitscore cutoff superior or equal to 50. For the postanalysis and genus-taxa encoding level estimations, an identity cutoff of 90% was applied to select reads assigned to a given genome. The read coverage of a gene was defined as the sum of the aligned length of each read respecting these cutoffs, expressed in nucleotide.

The peptide recruitment was based on the Mascot score reported by the Scaffold software when searching the database DB3 as for the DNA recruitment; no other cutoff was applied to identify peptides assigned to a reference genome. To compare relative expression between genes (*mxaF* and *xoxF*) (see Fig. S5), we defined the expression level of a given gene using the following calculation: (number_of_spectra/gene_length)/read_coverage. The relative expression ratio of 2 genes is the ratio of their gene relative expression.

1. Weisburg W, Barns G, Dale SM, Pelletier DA, Lane DJ (1991) 16S ribosomal DNA amplification for phylogenetic study. *J Bacteriol* 2:697–703.
2. Ochsenreiter T, Selezi D, Quaiser A, Bonch-Osmolovskaya L, Schleper C (2003) Diversity and abundance of Crenarchaeota in terrestrial habitats studied by 16S RNA surveys and real time PCR. *Environ Microbiol* 5:787–797.
3. Pruesse E, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35:7188–7196.
4. Yarza P, et al. (2008) The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst Appl Microbiol* 31:241–250.
5. Henckel T, Friedrich M, Conrad R (1999) Molecular analyses of the methane-oxidizing microbial community in rice field soil by targeting the genes of the 16S rRNA, particulate methane monooxygenase, and methanol dehydrogenase. *Appl Environ Microbiol* 65:1980–1990.
6. Green SJ, Minz D (2005) Suicide polymerase endonuclease restriction, a novel technique for enhancing PCR amplification of minor DNA templates. *Appl Environ Microbiol* 71:4721–4727.
7. Finn RD, et al. (2008) The Pfam protein families database. *Nucleic Acids Res* 36:D281–D288.
8. Olsen JV, et al. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol Cell Proteomics* 4:2010–2021.
9. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20:3551–3567.
10. Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20:1466–1467.
11. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 74:5383–5392.
12. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 75:4646–4658.
13. Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 4:207–214.
14. VerBerkmoes NC, Denef VJ, Hettich RL, Banfield JF (2009) Systems biology: Functional analysis of natural microbial consortia using community proteomics. *Nat Rev Microbiol* 7:196–205.
15. Zhang B, et al. (2006) Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Proteome Res* 5:2909–2918.
16. Tringe SG, et al. (2005) Comparative metagenomics of microbial communities. *Science* 308:554–557.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.

a

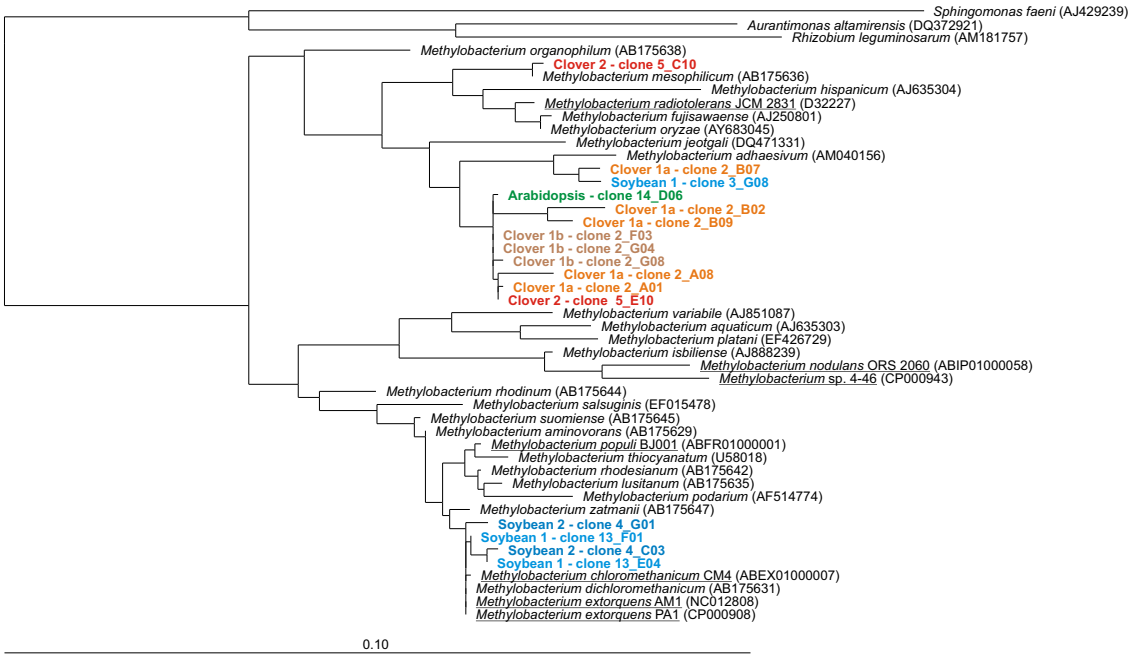


Fig. S1. Phylogenetic trees of 16S rRNA gene sequences obtained from clone libraries of all 6 samples. Phylogenetic relationship of nearly full-length 16S rRNA gene sequences to sequences of type strains and genome-sequenced strains (*underlined*) of the genera (*a*) *Methylobacterium*, (*b*) *Sphingomonas*, and (*c*) *Pseudomonas*. (*d*) The phylogenetic position of all other sequences detected in the clone libraries is shown with regard to the most closely related sequence and to sequences of cultivated reference organisms. All trees were constructed based on 1,388 nucleotide positions with the maximum likelihood algorithm PhyML. The bar represents 10% sequence divergence.

Fig. S1. Continued.

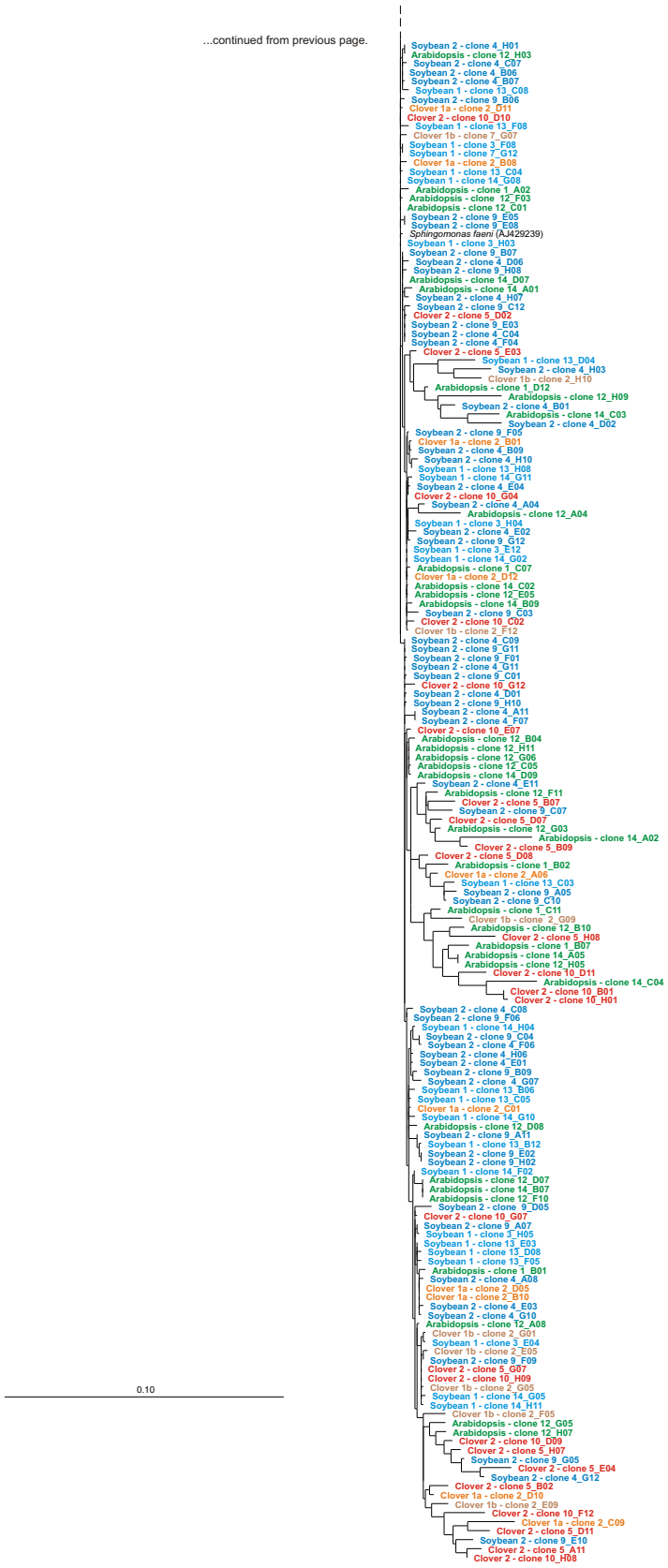


Fig. S1. Continued.

Fig. S1. Continued.

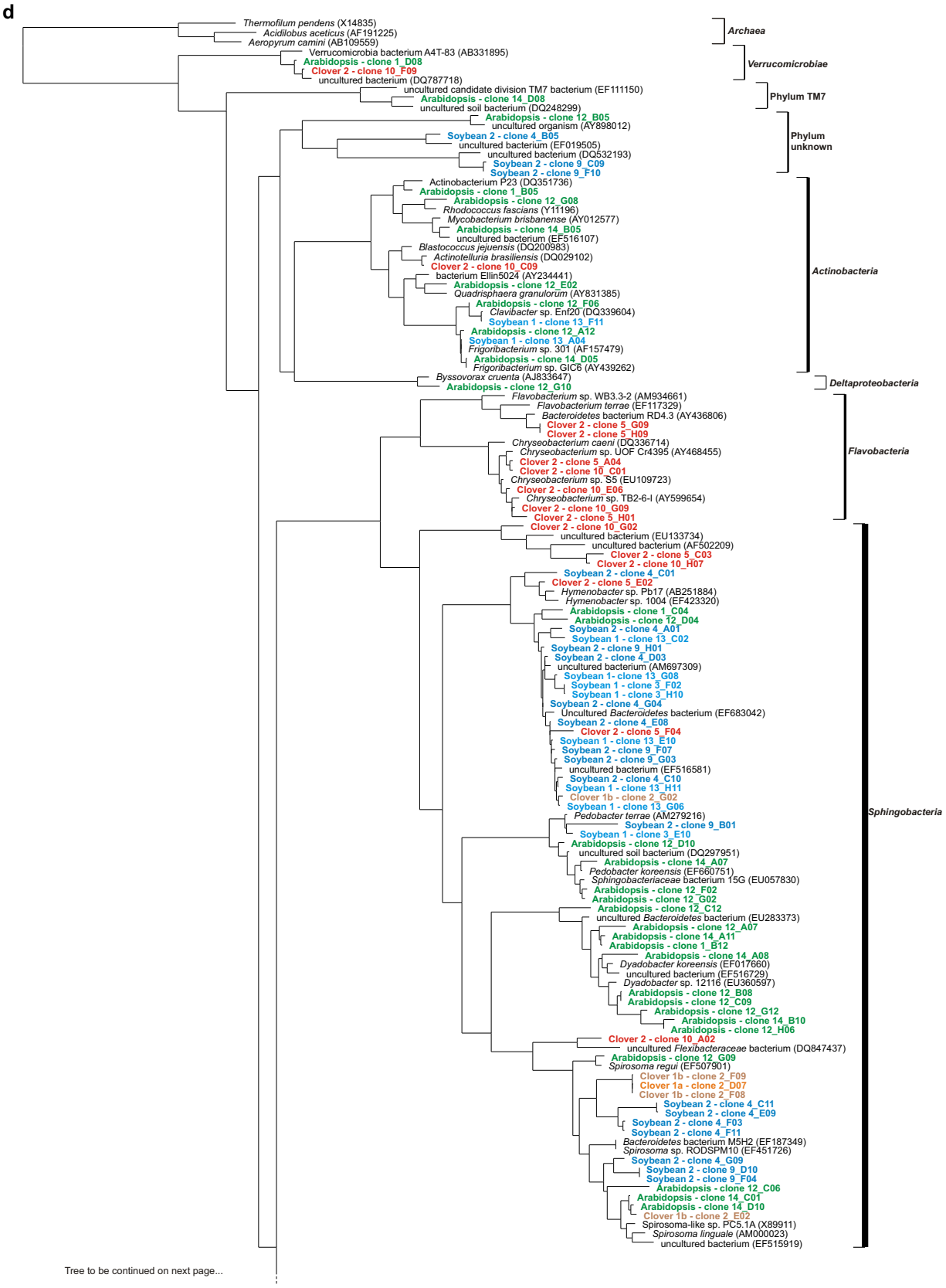


Fig. S1. Continued.

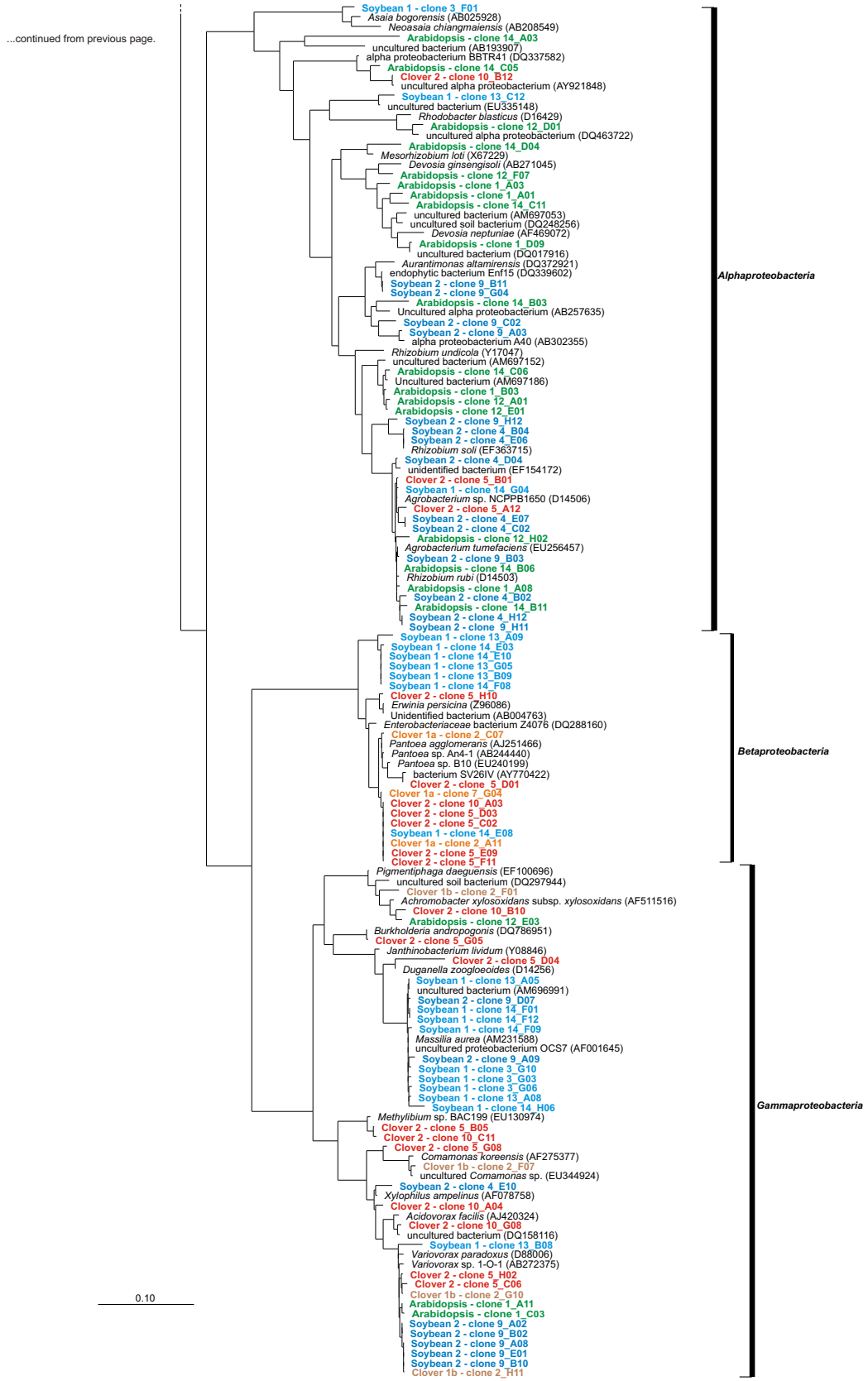


Fig. S1. Continued.

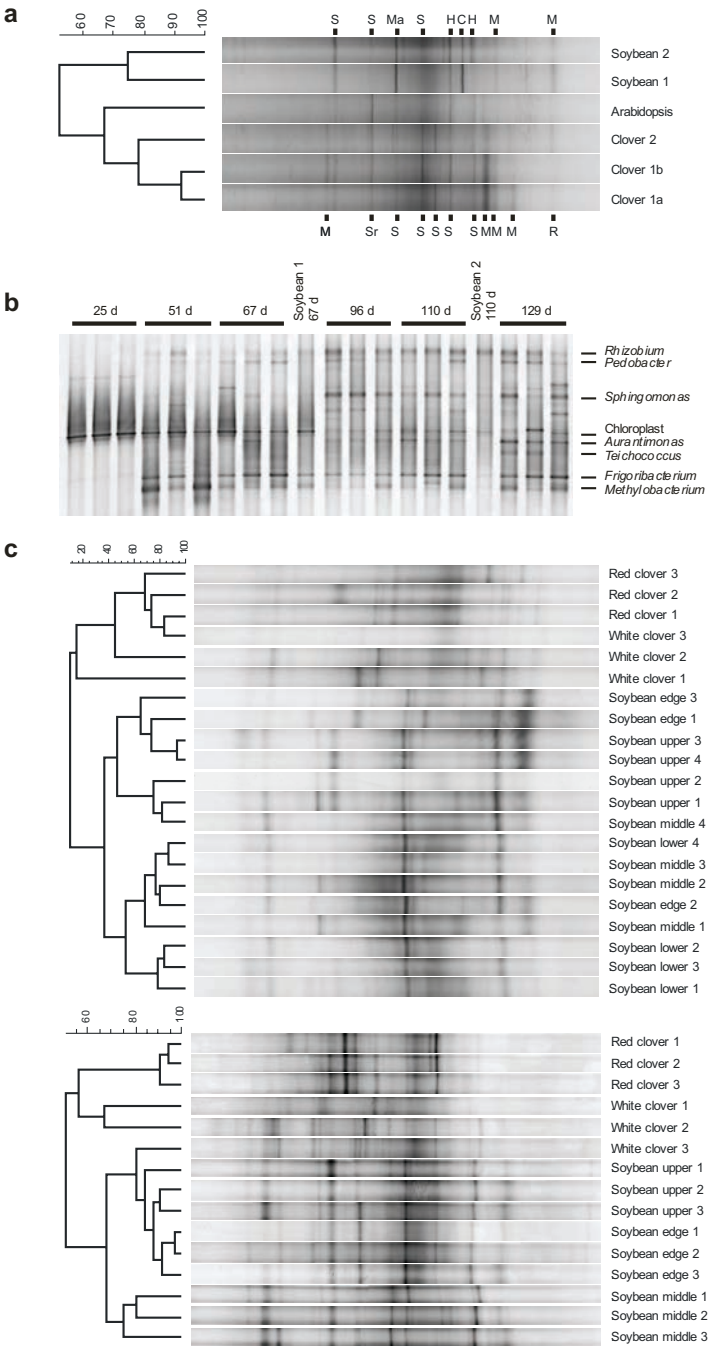


Fig. S2. Composition, similarities, and dissimilarities of microbial phyllosphere communities analyzed by denaturing gradient gel electrophoresis of 16S rRNA gene-amplification products. Cluster analyses of DGGE profiles were done using the unweighted-pair group method using arithmetic averages algorithm based on Pearson correlation coefficients. (*a*) Community composition of the 6 phyllosphere samples as seen based on DGGE banding patterns and sequencing of representative bands. The identity of bands from soybean and clover samples is indicated as follows: (H) *Hymenobacter*, (M) *Methylobacterium*, (Ma) *Massilia*, (R) *Rhodococcus*, (S) *Sphingomonas*, (Sr) *Spirosoma*, (C) Chloroplast. (*b*) Temporal succession of the bacterial phyllosphere community on leaves of soybean plants. Young leaves were collected at the beginning, while older leaves from the middle part of the plant were collected from day 67 on. The fingerprints suggest that soybean leaves were colonized by a rather stable bacterial community over the growing period. Early colonizers were still detectable at later developmental stages. Samples collected for the proteogenomic analysis ($n > 300$ leaves) reflect the community composition of the subsamples ($n = 3$ to 4 leaves per replicate) taken at the same time. (*c*) Comparison of microbial communities on soybean leaves taken from the upper, middle, and lower parts of the plant and, therefore, of different age. Additional samples were included from soybean grown at the edge of the field (from the middle part of the plant) and adjacent clover plants (white and red clover). The analysis was performed on material taken on days on which samples for metaproteomic analyses were taken: for example, day 67 (*Upper*) and day 110 (*Lower*). Bacterial communities on leaves of the soybean plants were in general more similar to each other than to communities on the clover plants, suggesting a plant species-specific phyllosphere colonization. Most of the dominant bands were detected on all different soybean samples, regardless of leaf age, leaf position, or plant position in the field (center versus edge), suggesting the presence of a spatially homogeneous bacterial phyllosphere population.

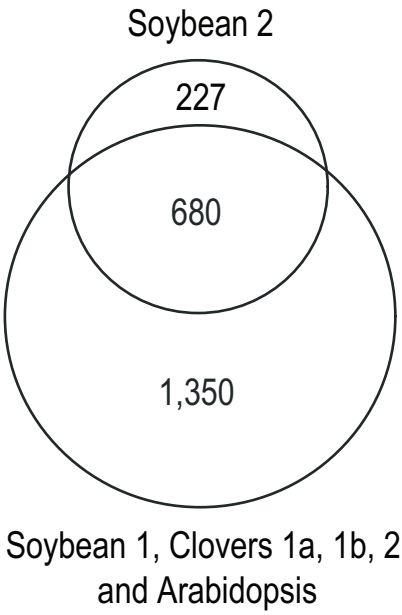


Fig. S3. Venn diagram showing the overlap of proteins identified in the Soybean 2 sample relative to the other plant phyllosphere samples. Peak lists of each sample were searched against DB2 and validated with Scaffold. The 6 resulting Scaffold files were merged together and only proteins assigned to bacteria were displayed.

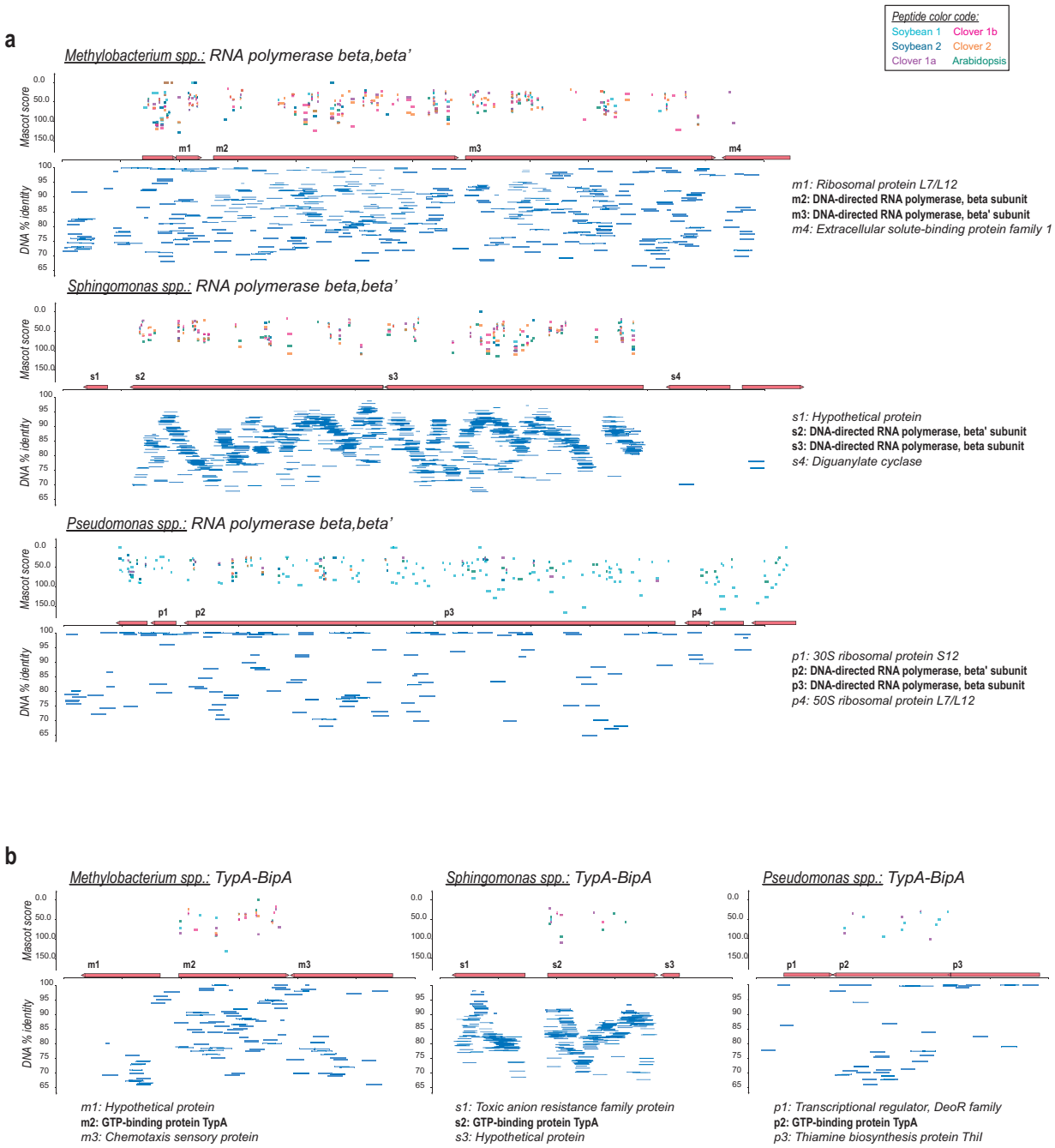
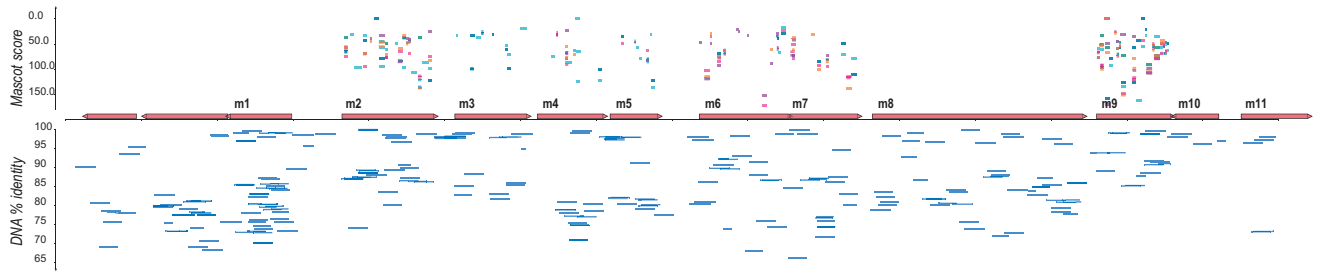


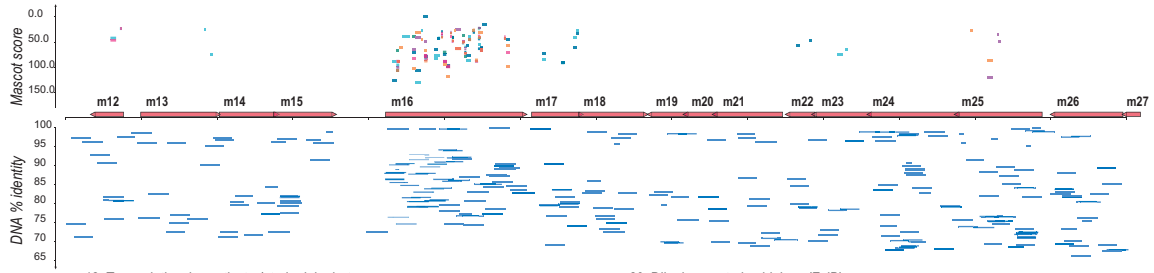
Fig. S4. Diverse fragment recruitment plots for proteins and functional clusters of particular interest. Both metagenomic (*Below x*-axis) and metaproteomic (*Above x*-axis) data are recruited on a reference genome. The color code indicates the sample source of a given peptide. (*a*) Exemplarily for a housekeeping protein, RNA polymerase beta and beta′ within the reference genera *Methylobacterium*, *Sphingomonas*, and *Pseudomonas* is shown. (*b*) GTP-binding protein TypA/BipA within the 3 reference genera is shown. This protein belongs to the ribosome-binding family of GTPases and is widely distributed in bacteria and plants. It has been shown to regulate multiple cell surface and virulence-associated components in *Escherichia coli* (1), and it is required for growth at low temperatures in Enterobacteria and *Sinorhizobium meliloti* (2, 3). (*c*) A major methylotrophy cluster in *Methyloacterium* containing genes for tetrahydromethanopterin-linked 1-carbon dissimilation and the serine cycle involved in 1-carbon assimilation, as well as genes for cofactor biosynthesis of tetrahydromethanopterin and pyrrolo quinoline quinone (4). Please note that *xoxF* (Fig. S5) is also encoded in this genomic region. (*d*) Motility cluster of *Pseudomonas*, covering ≈67kb. (*e*) The gene-annotated Beta-Ig-H3/fasciclin from *M. extorquens* PA1 attracted both DNA reads and peptides. 1. Farris M, Grant A, Richardson TB, O'Connor CD (1998) BipA: a tyrosine-phosphorylated GTPase that mediates interactions between enteropathogenic *Escherichia coli* (EPEC) and epithelial cells. *Mol Microbiol* 28:265–279. 2. Kiss E, Huguet T, Poinsot V, Batut J (2004) The *typA* gene is required for stress adaptation as well as for symbiosis of *Sinorhizobium meliloti* 1021 with certain *Medicago truncatula* lines. *Mol Plant Microbe Interact* 17:235–244. 3. Pfennig PL, Flower AM (2001) BipA is required for growth of *Escherichia coli* K12 at low temperature. *Mol Genet Genomics* 266:313–317. 4. Chistoserdova L, Chen SW, Lapidus A, Lidstrom ME (2003) Methylotrophy in *Methylobacterium extorquens* AM1 from a genomic point of view. *J Bacteriol* 185:2980–2987.

C

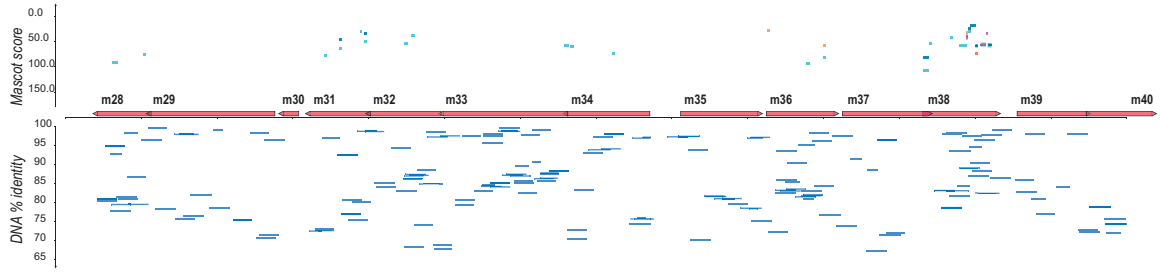
Methylobacterium spp.: Methylotrophy cluster



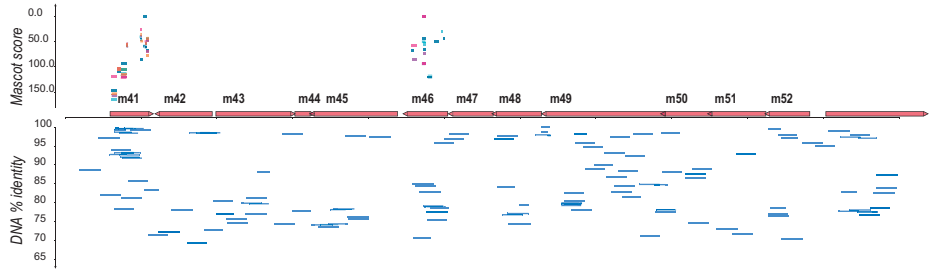
m1: Metallophosphoesterase
m2: Serine-glyoxylate aminotransferase (SGAT)
m3: Hydroxypyruvate reductase (HprA)
m4: Methylene tetrahydromethanopterin/tetrahydrofolate dehydrogenase (MtdA)
m5: Methenyl tetrahydrofolate cyclohydrolase (Fch)
m6: Malate thiokinase, large subunit (MtkA)
m7: Malate thiokinase, small subunit (MtkB)
m8: Phosphoenolpyruvate carboxylase (PpC)
m9: Malyl-CoA lyase (Mcl)
m10: Hypothetical protein
m11: Carboxymethylenebutenolidase



m12: Transcriptional coactivator/pterin dehydratase
m13: Putative ABC transporter periplasmic solute-binding protein
m14: Binding-protein-dependent transport systems inner membrane component
m15: ABC transporter related
m16: Homolog of PQQ-dependent methanol dehydrogenase large subunit (XoxF)
m17: Cytochrome C (XoxG)
m18: Conserved exported of unknown function (XoxJ)
m19: 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase (FolK)
m20: Dihydroneopterin aldolase (FolB)
m21: Dihydropteroate pyrophosphorylase (FolP)
m22: Protein of unknown function
m23: Protein of unknown function
m24: Pyrroloquinoline quinone (PQQ) biosynthesis protein E (PqqE)
m25: Bifunctional coenzyme pyrroloquinoline quinone (PQQ) biosynthesis protein C/D (PqqC/D)
m26: Pyrroloquinoline quinone (PQQ) biosynthesis protein B (PqqB)
m27: Pyrroloquinoline quinone (PQQ) biosynthesis protein A (PqqA)



m28: Response regulator MxbM
m29: Sensor kinase (MxbD)
m30: Protein of unknown function
m31: Formyltransferase/hydrolase complex Fhc subunit C (FhcC)
m32: Formyltransferase/hydrolase complex subunit D (FhcD)
m33: Formyltransferase/hydrolase complex Fhc subunit A (FhcA)
m34: Formyltransferase/hydrolase complex Fhc subunit B (FhcB)
m35: Ribofuranosylaminobenzene 5'-phosphate (RFAP) synthase (H4MPT biosynthesis) (MptG)
m36: Methylene tetrahydromethanopterin dehydrogenase (MtdB)
m37: OrfY
m38: Methenyl tetrahydromethanopterin cyclohydrolase (Mch)
m39: Orf5; m40: Orf7 (cofactor biosynthesis)

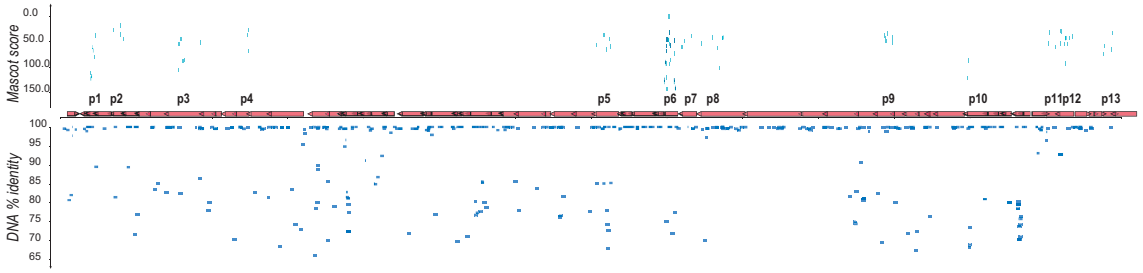


m41: Formaldehyde activating enzyme (Fae)
m42: Orf17 (Cofactor biosynthesis)
m43: Orf9 (Cofactor biosynthesis)
m44: Protein of unknown function
m45: Protein of unknown function
m46: Protein of unknown function
m47: Protein of unknown function
m48: Orf19 (cofactor biosynthesis)
m49: Orf20 (cofactor biosynthesis)
m50: Orf21 (cofactor biosynthesis)
m51: Orf22 (cofactor biosynthesis)
m52: Protein of unknown function

Fig. S4. Continued.

d

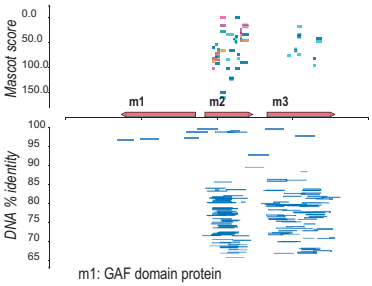
Pseudomonas spp.: Motility cluster *



p1 (71734681): Chemotaxis protein CheW
p2 (71737980): ParA family protein
p3 (71734503): Chemotaxis sensor histidine kinase CheA
p4 (71736626): Flagellar synthesis regulator FleN
p5 (71734449): Flagellar regulator FleQ
p6 (71733403): Flagellin
p7 (71737685): 3-oxoacyl-(acyl-carrier-protein) synthase III, putative
p8 (71735871): Glycosyl transferase, group 2 family protein
p9 (71735315): Flagellar basal body P-ring protein
p10 (71736953): Flagellar hook protein FlgE
p11 (71737331): Chemotaxis protein methyltransferase CheR
p12 (71734029): Chemotaxis protein CheV, putative
p13 (71735292): Hypothetical protein
** here only proteins (RefSeq PID between brackets) detected are reported.*

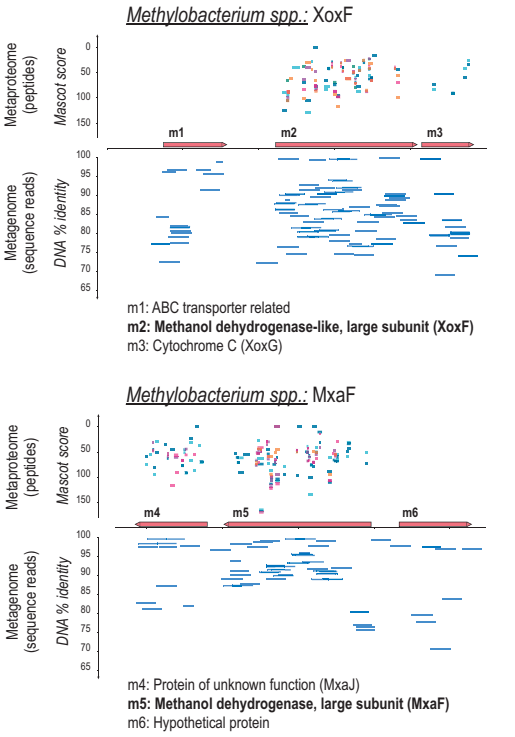
e

Methylobacterium spp.: Fasciclin



m1: GAF domain protein
m2: Beta-Ig-H3/fasciclin
m3: Hypothetical protein

Fig. S4. Continued.



protein name	*gene length (a.a.)*	*# reads identity >90%*	*read coverage (nt)*	*# peptides*	*# spectra*
XoxF	601	23	5,843	91	97
MxaF	626	44	10,912	368	407
Relative expression ratio computed for Soybean 2: XoxF/MxaF = 0.5 Relative expression ratio computed for all samples: XoxF/MxaF = 1.4					

Fig. S5. Fragment recruitment plots of MxaF and XoxF. Recruitment of both metagenomic (*Below x*-axis) and metaproteomic (*Above x*-axis) data on a *Methylobacterium* reference genome (*M. extorquens* PA1). Relative expression levels of *mxaF* and *xoxF* were calculated to compare expression strength of both genes. The color code indicates the sample source of a given peptide.

Other Supporting Information Files

Table S1
Table S2
Table S3
Table S4
Table S5
Dataset S1

6 MICROBIAL COEXISTENCE AND GENOME EVOLUTION

6.1 PREFACE

Environmental preferences and ecological interdependencies of microorganisms are poorly characterized and remain difficult to assess. Today, the vast amount of available 16S rRNA sequences, sampled from various environments, allows us to globally investigate patterns of co-occurrence among microbes and structures of natural microbial communities. Moreover, the continuous acquisition of whole-genome sequence data from various organisms enables comparative genomics within an ecological framework. In the following study, we developed a method to connect the molecular information contained in the genome of a lineage to the co-occurrence patterns of that lineage around the globe.

For this research project[18], I developed and implemented in totality the analytical pipeline (except for the third party softwares included in the workflow) to globally detect preferential coexistences among microorganisms. I also performed the comparative genomics analyses in order to gain insights into the shared functional capabilities of coexisting microbes. I created all figures of the publication except figures 3, S1 and S5.

6.2 A GLOBAL NETWORK OF COEXISTING MICROBES

The publication is included below.

Research

A global network of coexisting microbes from environmental and whole-genome sequence data

Samuel Chaffron,[1] Hubert Rehrauer,[2] Jakob Pernthaler,[3] and Christian von Mering[1,4]

[1]*Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, CH-8057 Zürich, Switzerland;* [2]*Functional Genomics Center Zurich, University of Zürich and ETH Zürich, CH-8057 Zürich, Switzerland;* [3]*Limnological Station of the Institute of Plant Biology, University of Zürich, CH-8802 Kilchberg, Switzerland*

Microbes are the most abundant and diverse organisms on Earth. In contrast to macroscopic organisms, their environmental preferences and ecological interdependencies remain difficult to assess, requiring laborious molecular surveys at diverse sampling sites. Here, we present a global meta-analysis of previously sampled microbial lineages in the environment. We grouped publicly available 16S ribosomal RNA sequences into operational taxonomic units at various levels of resolution and systematically searched these for co-occurrence across environments. Naturally occurring microbes, indeed, exhibited numerous, significant interlineage associations. These ranged from relatively specific groupings encompassing only a few lineages, to larger assemblages of microbes with shared habitat preferences. Many of the coexisting lineages were phylogenetically closely related, but a significant number of distant associations were observed as well. The increased availability of completely sequenced genomes allowed us, for the first time, to search for genomic correlates of such ecological associations. Genomes from coexisting microbes tended to be more similar than expected by chance, both with respect to pathway content and genome size, and outliers from these trends are discussed. We hypothesize that groupings of lineages are often ancient, and that they may have significantly impacted on genome evolution.

[Supplemental material is available online at http://www.genome.org.]

Symbiosis—as defined in its broadest sense (de Bary 1879; Saffo 1993)—is widespread in nature, ranging from obligatory mutualistic partnerships to commensalism to clearly detrimental, parasitic interactions (Paracer and Ahmadjian 2000). The phenomenon is not restricted to a particular domain of life, but can occur, for instance, between bacteria, archaea, and protists, which, in turn, can live together inside a specific animal host (Brauman et al. 1992; Tokura et al. 2000). Many instances of symbiosis are known, but they are not always understood mechanistically. The situation may not always be stable either: Symbionts may "cheat," and/or compete among each other for a third partner (Palmer et al. 2003; Ferriere et al. 2007; Johnstone and Bshary 2008).

Leaving aside macroscopic organisms, symbiosis and local coexistence among single-celled microbes are even less well characterized. The extent, specificity, and stability of microbial associations are difficult to assess systematically in the environment, since elaborate staining procedures and/or molecular sequencing are needed in order to detect and differentiate between microbial lineages in situ. Nevertheless, several close partnerships between microbial species have already been identified. These include consortia of methane-oxidizing archaea and sulfate-reducing bacteria (AOM, "anaerobic oxidation of methane") (Boetius et al. 2000; Caldwell et al. 2008; Knittel and Boetius 2009); consortia of phototrophic green sulfur bacteria surrounding motile beta-proteobacteria (Overmann and Schubert 2002; Wanner et al. 2008); consortia of sulfate reducers, sulfate oxidizers, and other lineages inside marine, gutless oligochaete worms (Dubilier et al. 2001; Woyke et al. 2006; Ruehland et al. 2008); and consortia of extremophilic lineages conducting ferrous iron oxidation in acidic pyrite mine run-offs (Tyson et al. 2004). Such groupings probably do not constitute "symbiosis" in a classical sense (Saffo 1993), but they are typically interpreted as syntrophic associations in which one partner consumes metabolites produced by the other. In addition, predatory and parasitic relationships are also known. An example for the latter is *Nanoarchaeum equitans*, a small archaeon that appears to be an obligate parasite of another archaeal species (Huber et al. 2002; Forterre et al. 2009). Despite such specific findings, the discovery of microbial associations has so far been largely interest-driven (or even fortuitous), meaning that a comprehensive picture of microbial coexistence has yet to emerge.

The notion that microbes in the environment perhaps exist in a less solitary manner than commonly assumed is also supported by the rapidly accumulating knowledge on intra- and interspecies microbial communication (Ryan and Dow 2008; Shank and Kolter 2009). Essential activities of single species such as nutrient uptake, biofilm formation, or cellular differentiation can be organized and synchronized by communication and cooperation (Parsek and Greenberg 2005; Waters and Bassler 2005; Kolter and Greenberg 2006; Gibbs et al. 2008; Ng and Bassler 2009). While it is less clear whether and to what extent microbes may interact with other species via specific communication, some bacteria are known to "eavesdrop" and to even respond to signals that they cannot themselves generate (Visick and Fuqua 2005). In addition, an interspecies relationship has been shown to evolve and quickly deepen in a laboratory evolution experiment (Hansen et al. 2007; Harcombe 2010).

Apart from the few cases of well-described, specific interactions, relatively little is known about how natural microbial assemblages form and how they are structured, if at all (Ruan et al. 2006; Horner-Devine et al. 2007; Fuhrman and Steele 2008; Raes and Bork 2008; Fuhrman 2009). They are often taxonomically highly complex and can encompass hundreds of different species, and at least some aspects of the composition of any given community are thought to be based on historical contingency (Martiny et al. 2006). Moreover, naturally occurring communities are difficult

[4]Corresponding author.
E-mail mering@imls.uzh.ch.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.104521.109. Freely available online through the *Genome Research* Open Access option.

to reassemble and/or study under controlled conditions in the laboratory, since most of the constituting lineages are not available in pure cultures (typically more than 95% of species present in a given sample cannot be cultivated) (Staley and Konopka 1985). The difficulties in cultivating microbes have often been linked to slow growth and unknown nutritional requirements, but might also be partly attributed to their synecology—for example, reflecting a need to coexist within a biofilm or to aggregate together with partner species in order to ameliorate adhesion (Min and Rickard 2009) or to dispose of otherwise inhibitory metabolic products.

Since the establishment of the first comprehensive microbial phylogeny using the 16S rRNA gene (Fox et al. 1980) and the invention of techniques for rapidly generating large blocks of 16S rRNA sequence data (Lane et al. 1985; Giovannoni et al. 1990; Ward et al. 1990), a great variety of environments have been sampled to study microbial diversity in situ. Today, the 16S rRNA gene remains the marker of choice for identifying microbes in their environments, and the size of databases dedicated to this gene is growing exponentially (Desantis et al. 2006; Pruesse et al. 2007; Cole et al. 2009). In addition, environmental sequences are increasingly being annotated with contextual information (e.g., geographic position, temperature). An important effort to define and standardize such sequence meta-data has been initiated within the Genomic Standards Consortium (Field et al. 2008a,b), specifically by developing the MIENS standard (minimum information about an environmental sequence). However, the existing annotations of many of the legacy sequences in the databases will have to be migrated to such standards, which requires considerable effort. The results of such efforts are increasingly being made available in integrated resources such as CAMERA (Seshadri et al. 2007), IMG/M (Markowitz et al. 2008), and megDB (Kottmann et al. 2010), but at present only a small minority of 16S rRNA sequences have geo-referencing or other contextual information.

Using 16S rRNA sequences in combination with other data, classical ecological questions including species (co)-occurrence and diversity have also been addressed extensively in microbes (Bell et al. 2005; Langenheder et al. 2006; Ruan et al. 2006; Horner-Devine et al. 2007; Smith 2007; Langenheder and Prosser 2008). In doing so, many of the concepts that have originally been developed for macroscopic organisms have been adapted and applied to microbes. However, these studies have mostly focused on one specific environment, or one specific lineage, at a time (e.g., Alonso et al. 2007; Newton et al. 2007; Fuhrman and Steele 2008) (this way, ecological questions can be studied in a more defined setup). What has not been addressed much, so far, is the global partitioning of microbial lineages among all sampled environments. Here, we take a first step in this direction, by systematically studying a current snapshot of the complete data set of full-length 16S rRNA sequences. We search for groups of lineages that occur together more often than expected by chance, and we connect this information to genomic data, as well as to the limited metadata that are available regarding the sampling sites (the latter information stems mostly from free-text annotations provided at the time of database submission). We find that the assortment of lineages and environments is clearly nonrandom, and that specific and recurring associations among lineages can be described, at various levels of detail and phylogenetic resolution.

Results and Discussion

In order to comprehensively characterize the occurrence of microbial lineages in the environment, we first grouped publicly available, full-length 16S rRNA sequences at various levels of sequence identity, thereby creating unsupervised sets of "operational taxonomic units" (OTUs; see Methods for details). Each OTU was assigned a taxonomic annotation that reflected the consensus of its member sequences, and a single sequence was chosen to represent each OTU in subsequent sequence comparisons. Next, we comprehensively compiled environmental "sampling events" of 16S sequences; such an event is defined here as a unique combination of submitting authors, project title, and isolation source, as annotated in the respective database records. We assumed that sequence entries for which all three fields are exactly identical were sampled together, at a given site. Our procedure (Fig. 1) thus resulted in a large matrix that connects OTUs to environmental sampling events (Table 1). Depending on the OTU definition, this matrix contained roughly between 700 and 5000 distinct OTUs, which were mapped to roughly 3000 distinct sampling events (we only retained sampling events that encompassed at least two OTUs, and conversely, only OTUs that were observed in at least three sampling events).

Next, we examined this matrix for any non-random assortment of OTUs to environments, which would manifest itself as groups of OTUs observed together more often than expected by chance. Our underlying null model is that of global, random dispersal of lineages across environments (Harvey et al. 1983; Finlay 2002; Kunin et al. 2008a; Hubert et al. 2009), and essentially corresponds to the first part of Baas Becking's enigmatic statement, "Everything is everywhere, but, the environment selects" (de Wit and Bouvier 2006). While this null model is clearly not applicable for macroscopic organisms with distinct biogeographic distribution patterns, it does represent the simplest default assumption for microbes, and it is appropriate for the very large geographical and temporal scales that we consider here. By computing the hypergeometric probability of pairwise co-occurrences and correcting for multiple testing, we found that, indeed, a large number of statistically significant associations between OTUs can be observed, irrespective of the precise choice of OTU definition cutoff (Fig. 2; Supplemental Fig. S1; Table 1). A concrete example for such an association is shown in Figure 1B (data from Sorensen et al. 2005; Baati et al. 2008; Isenbarger et al. 2008; Sahl et al. 2008; R Amdouni, E Ammar, H Baati, N Gharsallah, and A Sghir, unpubl.): a well-characterized lineage of *Cyanobacteria* (belonging to the halophilic *Euhalothece*) (Garcia-Pichel et al. 1998) was observed to be associated with an uncharacterized lineage having no cultivated or named representatives (a monophyletic sister group of the *Psychroflexus* lineage [*Bacteroidetes*]). This particular association is based on three independent sampling events in which both lineages had been observed together, by three distinct laboratories in three distinct countries. Considering that the OTU definition in this case is relatively narrow (97%) and that this association occurs against a backdrop of about 2800 sampling events covering more than 5000 OTUs, the observation becomes highly significant ($P < 3 \times 10^{-06}$; after multiple testing adjustment). Overall, several thousand of such associations could be identified. To assess the effects of potential biases in the sampling data, and in order to estimate our false discovery rate (FDR) empirically, we performed a conservative randomization of our data—by keeping constant the size distributions of both sampling events and OTUs, but shuffling the connections between OTUs and sampling sites. This resulted in a reduction of the number of reported associations by >99% for most of the OTU definition cutoffs (Table 1), which translates to FDRs of ~1%, except at very broad OTU definitions (i.e., when setting the OTU clustering cutoff to 85% sequence

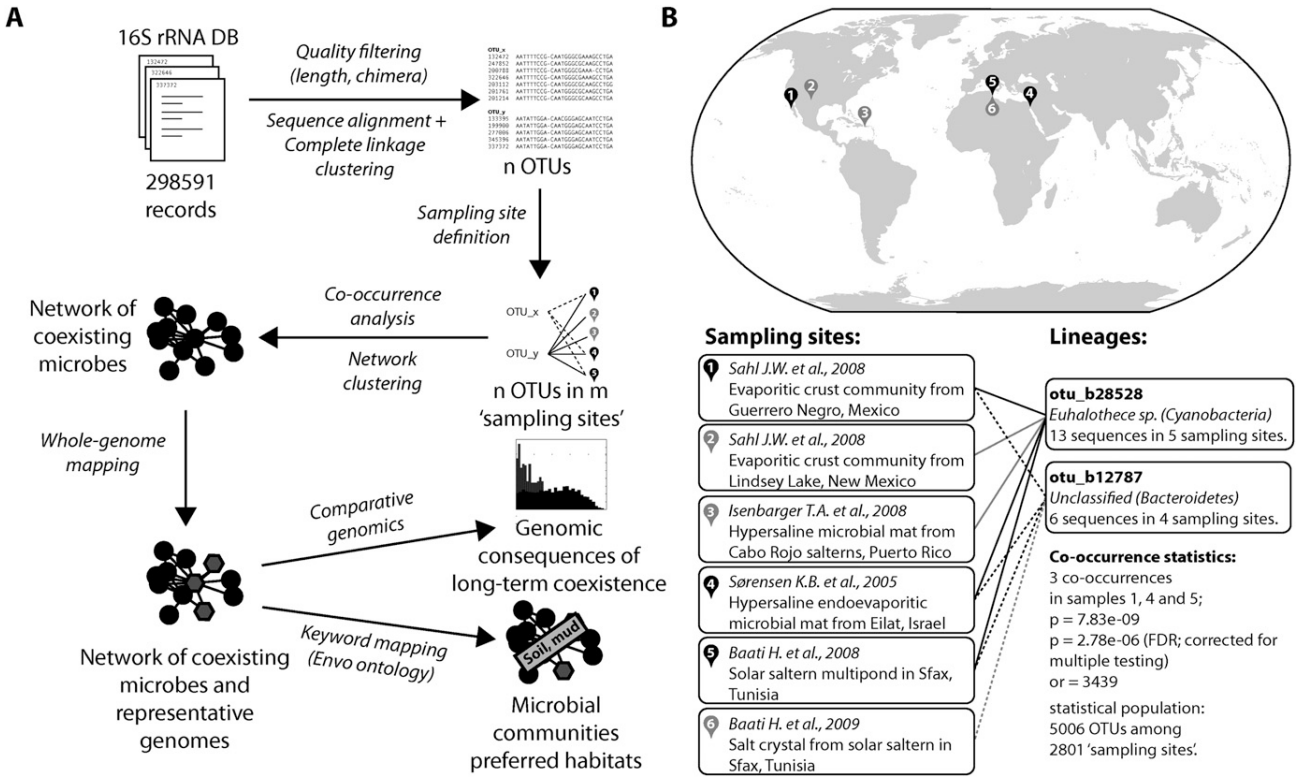


Figure 1. Detection of coexisting microbial lineages. (*A*) Schematic description of the analysis procedure. Publicly available 16S ribosomal RNA sequences are first grouped into operational taxonomic units (OTUs), then annotated according to unique environmental sampling events, and finally searched for statistically significant co-occurrences. Where available, completely sequenced genomes are mapped onto the resulting network, which is then clustered and annotated. (*B*) Example for a specific lineage association. The two lineages (defined at a 97% 16S sequence identity cutoff) have been sampled overall relatively rarely, but they occurred together three times, at three distinct sites. (or) Odds ratio. Under "Sampling sites," the investigative work of "Baati H. et al., 2009" refers to R Amdouni, E Ammar, H Baati, N Gharsallah, and A Sghir (unpubl.).

identity or less) (see Table 1). The few remaining false-positive associations were observed mainly among widely sampled lineages known to occur inside the mammalian digestive tract; this likely reflects the strong study bias toward 16S gene sequences of this habitat (Ley et al. 2008; Hamady and Knight 2009).

In addition to assessing the statistical significance, we also computed a "specificity" value (or "association strength") for all OTU pairs. This value corresponds to the Jaccard similarity; it is 1.0 if a pair of OTUs is always observed together (but never separately), and zero for a pair of OTUs that is always observed in distinct environments, but never together. Remarkably, we found associations at both extremes of specificity (i.e., close to 1.0 or close to zero) (see Supplemental Fig. S3 for the overall distribution). An example for the former is shown in Supplemental Figure S1A: a previously undescribed bacterial OTU (a sublineage of the candidate division *JS1*) was observed together with a specific *Methanosarcinales* lineage in marine sediments, again by three distinct laboratories (once in the Mediterranean, twice in the Gulf of Mexico at distinct sites). However, in this case, the two lineages were never found separately, in any of the 2800 environment samplings we studied. The likelihood of observing such a specific association by chance is again very low ($P = 1.1 \times 10^{-7}$ after correction for multiple testing). An example for a less specific but nevertheless highly significant association is shown in Supplemental Figure S1B: A lineage of gamma-Proteobacteria (nosocomial pathogens from the genus *Stenotrophomonas*) was frequently observed together with a lineage of *Bacilli* (genus *Staphylococcus*). The two lineages were sampled together 10 times—by seven distinct laboratories—in various air samples, skin samples, dust, and on Chinese cabbage. The association is highly significant ($P < 10^{-9}$ after correction for multiple testing), but less specific: Both lineages have also been observed separately (in 32 and 17 sampling sites, respectively). Not all of the latter observations were related to skin samples. *Stenotrophomonas*, for example, may also form distinct blooms in shallow coastal lagoons (Piccini et al. 2006). While co-occurrence alone cannot offer any mechanistic explanation for lineage associations, the additional information in the specificity of an association does provide a constraint when discussing possible scenarios (obligatory mutualism, for example, would be expected to result in a high association specificity). Barring any additional information, we did choose to interpret our observed associations conservatively, by assuming that they for the most part simply reflect shared or overlapping niche preferences. Instances of undescribed, specific mutualisms and parasitisms are presumably contained within our findings, but additional experimental follow-ups will be required for a detailed characterization of such interactions (Orphan 2009). That notwithstanding, this first part of our analysis already provides an empirical base for discovery and allows us to explore more specific hypotheses about the reasons for the coexistence of sets of uncultured genotypes.

Next, we searched our observed co-occurrence relations for previously known microbial associations (Supplemental Fig. S1).

Table 1. Overview of sampled microbial lineages at various levels of OTU definitions

OTU definition (%)	80	85	90	95	97	98	99
No. of OTUs	1059	3142	9018	25,142	38,186	48,144	65,807
No. of OTUs after filtering	713	1627	3286	5001	5006	4697	4228
No. of sampling sites after filtering	2698	2826	2918	2931	2801	2633	2312
No. of co-occurrence tests	25,3828	1,322,751	5,397,255	12,502,500	12,527,515	11,028,556	8,935,878
No. of coexisting OTU pairs (FDR = 0.001)	14,421	32,908	67,219	78,529	83,614	88,636	104,876
Random data: no. of coexisting OTU pairs (FDR = 0.001)	5618	3515	1006	693	503	834	433
FDR (estimated by permutations)	0.3896	0.1068	0.0150	0.0088	0.0060	0.0094	0.0041
No. of OTUs with mapped genome	NC	NC	NC	350	499	598	663
Coexisting genome pairs (FDR = 0.001)	NC	NC	NC	410	303	232	200

The table provides numerical details on the raw data and the results, and also illustrates the effects of changing the phylogenetic resolution at which the analysis is performed. For very narrowly defined OTUs, many lineages have to be discarded because they do not occur in a sufficiently large number of samples. Conversely, for very broadly defined OTUs, the statistical false discovery rate becomes too high, since many of the more abundant OTUs are seen to co-occur even after conservative randomization of sampling sites. NC, Not computed.

While we did not recover the known association between the *Nanoarchaeum* and *Ignecoccus* lineages, nor the *Chlorochromaticum* consortium, we did find strong evidence for AOM consortia (Supplemental Fig. S1D). We also observed the known association between the lineages *Leptospirillum* (phylum Nitrospira) and *Acidothibacillus* (phylum Proteobacteria), both of which are known to thrive in acidic bioleaching environments. In this case, the association we found was remarkably strong and specific: Out of 21 independent observations of *Leptospirillum* (by 18 distinct author teams in various settings), all but a single one also included observations of *Acidothiobacillus* (i.e., 20 out of 21; $P < 10^{-35}$) (Supplemental Fig. S1C) (in this case, the OTU clustering distance was 90%). Remarkably, this association appeared to be somewhat asymmetric: *Acidothiobacillus* did occur occasionally without its partner (in an additional 18 sampling events), suggesting that the mutual dependencies might not be equally strong in both directions. As a further test of our associations, we conducted an independent co-occurrence search of microbial lineages in the published literature (Supplemental Fig. S5). The frequencies of co-mentions of species names in PubMed can, indeed, reveal ecological associations (Freilich et al. 2010), albeit limited to those lineages that are already validly named and for which cultivated type strains typically exist. We find that more than 70 of our pairwise associations (counting nonredundantly at the genus level) can, indeed, be confirmed by the published literature, that is, their co-mention counts rise above a conservative randomization of species names and PubMed entries (Supplemental Fig. S5). Apart from the known associations, we also observed a large number of previously undescribed interactions, many of which involved unclassified lineages without any cultured or named representative (discussed below; the full set of associations is also available for browsing online). It should be noted that our data set likely misses some aspects of microbial coexistences, due to experimental biases in the generation of 16S rRNA sequences. In particular, the frequent choice of primers that will not target archaeal sequence types (Muyzer et al. 1995) in environmental studies may lead to an underestimation of the association between bacteria and archaea (but see Supplemental Fig. S1 and Fig. 5, below, for examples).

The observed associations were not limited to pairwise co-occurrences. When plotting the associations as a graph, a densely connected network of OTUs emerged (Fig. 2A). The topology of that network is clearly nonrandom; it exhibits a high clustering coefficient, short average minimum path length, and a connectivity degree distribution that has no characteristic maximum (i.e., the network is roughly matching the "scale free, small-world" criteria) (Barabasi and Oltvai 2004). This topology suggests that the network can be meaningfully partitioned, and that doing so should reveal modules of densely connected microbial lineages; these might be regarded as the microbial equivalents of the "syntaxa" of vegetation analysis. One such possible partitioning is shown in Figure 2C; it conveys more information than a simple list of pairwise co-occurrences because it groups specific lineages, at the exclusion of others. Module formation can occur even if the various pairwise correlations are not all highly significant (e.g., due to undersampling); this is because a certain fraction of missing or poorly scoring associations can be tolerated as long as the overall topology remains that of a tightly linked module. Furthermore, partitioning allows the annotation of keywords that describe the commonalities among the associated sampling sites of the various modules (Fig. 2C; see Methods). Among the modules, we observed intriguing cases where all or the majority of lineages have not been characterized before. An example is shown in Figure 3 (data from Heijs et al. 2005; Inagaki et al. 2006; Ley et al. 2006; Lloyd et al. 2006; Isenbarger et al. 2008; Li and Wang 2008; Li et al. 2008; Zhang et al. 2008; Harrison et al. 2009; Takeuchi et al. 2009; Ghosh et al. 2010)—five lineages that are co-occurring very specifically in certain marine sediments; they are from three distinct phyla, and each lineage is entirely uncharacterized. (A closer phylogenetic analysis revealed that the two *Planctomycetes* OTUs are related to each other, to the exclusion of other *Planctomycetes* lineages; they have been found also in other marine and freshwater environments, and our co-occurrence thus defines a more restricted home context for this lineage.) Specific modules such as this example are striking and likely provide a first glimpse onto hitherto undescribed microbial consortia.

While our 16S-based OTUs provide fairly objective coverage of phylogenetic lineage space, they do not, in themselves, contain any information about molecular and ecological functions. We therefore attempted to represent each OTU by its best match among completely sequenced genomes, to the extent that the latter are available (see Methods). Strains for which complete genomes have been sequenced do not usually originate from the environmental samplings described here. However, as long as they are closely related to the OTU in question, they may suffice to reveal broad genomic trends related to coexistence. The validity of this approach is based on two observations/assumptions. First, our co-occurrence analysis is already enriching for lineages that are

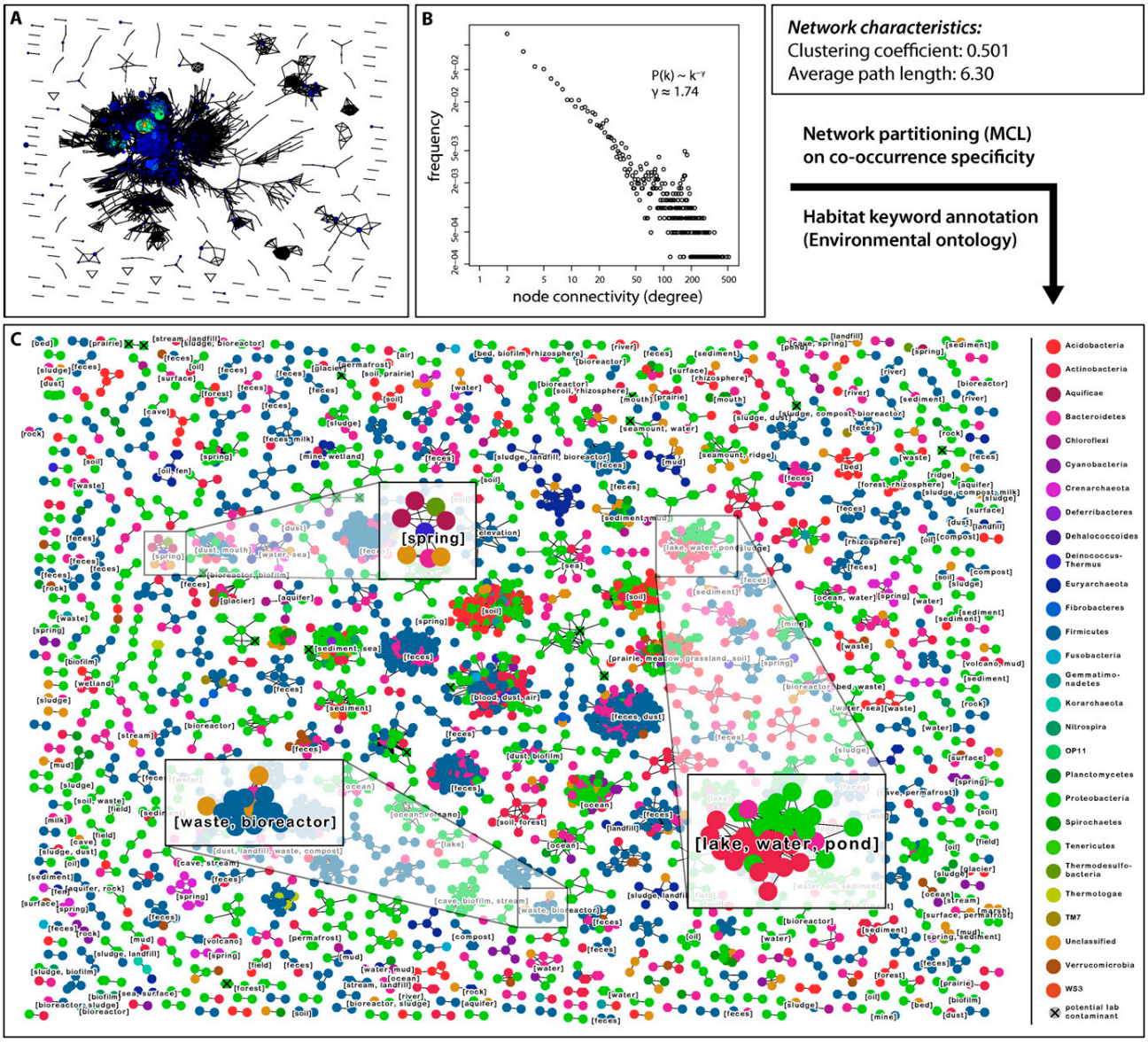


Figure 2. Global network of coexisting microbial lineages. (*A*) Overview of the network of lineage associations. Each node denotes a microbial lineage, and each line a significant co-occurrence relationship. Node size is proportional to the number of sequences in the lineage, and node color indicates the connectivity degree of a node (along a color gradient: blue, low connectivity; red, high connectivity). Throughout the figure, the OTU definition cutoff is at 97% sequence identity, and the *P*-value cutoff for an association is 0.001 (i.e., FDR after correction for multiple testing). (*B*) Connectivity degree distribution plot for the network in *A*. The distribution is coarsely compatible with a power law distribution. (*C*) Same network as in *A*, but partitioned using unsupervised Markov clustering, to reveal modules (clusters) of co-occurring lineages. Here, node color denotes taxonomic classification at the phylum level. Lineages suspected to contain potential laboratory contaminants (Tanner et al. 1998; Barton et al. 2006) are mainly observed in small clusters, and are marked with a small black X (17 such lineages in total).

likely abundant (Pedros-Alio 2006) and that can be widely found and easily accessed (each OTU had to be sampled at least three times to be included here). And, second, there seems to be a notable stability of environmental habitat preferences among microbial lineages in general (Von Mering et al. 2007; see also below). This suggests that a sequenced strain may represent other members of its OTU in terms of its genomic content and aspects of its ecology even if it has diverged from them to some degree. We were able to map between 350 and 660 genomes to a subset of our OTUs (this depends on the OTU resolution; notably it also means that a significant fraction of sequenced genomes currently cannot be connected to an OTU that has been repeatedly observed in the environment). This mapping translates to between 200 and 410 significant partnerships for which genomic information is available for both partners, covering a small but significant fraction of all the instances of co-occurrence we detected. To our knowledge, this is the first time that a global, environmentally motivated association network between genomes has been constructed.

We used this network to objectively assess potential constraints on genome evolution, which might be a consequence of

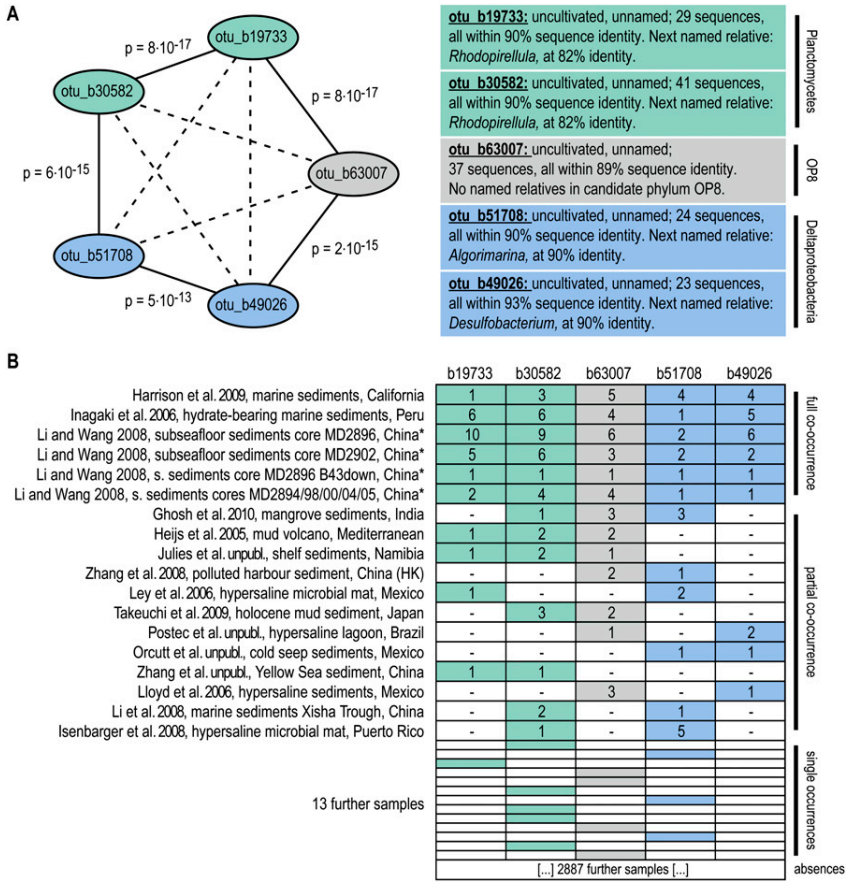


Figure 3. Example of a novel, previously undescribed module of coexisting lineages. (*A*) Five distinct microbial lineages are shown; they belong to three different phyla and are defined at an OTU-clustering distance of 90% sequence identity at the 16S rRNA gene. The five lineages have been exclusively observed through environmentally sampled sequences and have not been named. (*B*) The table shows all occurrence counts of these lineages among our sampling data; the *P*-values indicated have been corrected for multiple testing, against the background of all lineages defined at 90%. Adjusted *P*-values (FDR; p) and odds ratios (or) are indicated. (*) The samples by Li et al.(2008) have been collected at distinct sites, covering a distance of more than 600 miles; collection was at different water depths and sampling dates. Investigators involved in unpublished work are as follows: E Julies, V Bruechert, and BM Fuchs; B Orcutt, SB Joye, S Kleindienst, K Knittel, A Ramette, A Rietz, V Samarkin, T Treude, and A Boetius; A Postec, R Warthmann, C Vaconcelos, K Hanselmann, and J McKenzie; Z Zhang, H Xiao, and X Tang.

the association of a genome to its preferred environment and to other lineages in that environment. We observed four highly significant trends among co-occurring genomes: They tend (1) to have more similar genome sizes, (2) to be more similar in GC content (i.e., the fraction of the genome consisting of guanine and cytosine), (3) to be more similar with respect to relative coverage of functional pathways, and (4) to be phylogenetically more closely related than randomly selected pairs of genomes (Fig. 4). The latter trend was also visible from 16S sequences alone (Supplemental Fig. S2). The trend to phylogenetic relatedness is presumably the easiest to rationalize: Pairs of ecologically associated lineages, which are also closely related phylogenetically, would arise naturally assuming that neither lineage had changed their habitat preferences since they split from their last common ancestor. We indeed observe this signal and detect that it extends surprisingly far back in time: Lineages that have diverged up to 10% at the 16S sequence identity level are still clearly enriched among environmentally associated pairs (Fig. 4A; the peak seen at 15% sequence divergence is largely due to a single, well-covered cluster; see Supplemental Figs. S6, S8). In principle, this relatedness signal could also explain our three other observations: Phylogenetically related genomes are known to exhibit similar GC contents, genome sizes, and functional composition. To assess this possibility, we tested these three signals for independence from the phylogenetic signal, by correcting for the underlying correlations as learned from randomly selected genome pairs (Fig. 4). In the case of GC content similarity, we find that the signal can, indeed, be largely explained by phylogenetic relatedness alone—it is not an independent observation. This would argue against environmental selection on GC content, at least at longer time scales, and it gives further support to algorithms that partition environmental sequences based on genomic signatures (McHardy and Rigoutsos 2007; Mrazek 2009). In contrast, importantly, we observed that both genome size similarity and functional similarity could not be explained solely by phylogenetic relatedness. For example, while randomly selected pairs of genomes have genome sizes that can vary considerably, environmentally associated genome pairs tend to level off at ~20%–30% genome size difference, on average (Fig. 4F, $P < 10^{-13}$). This is remarkable because it suggests that a given environment tends to select for a particular optimal genome size range, even across distinct lineages; furthermore, it suggests that lineages spend sufficient time in their preferred environments to allow for these optimal genome sizes to be selected for and maintained (against a mutational spectrum that is thought to be largely biased toward deletions in bacteria) (Mira et al. 2001; Nilsson et al. 2005). Our observation confirms what has been known anecdotally from a number of environments: Planktonic marine environments, for example, persistently select for small to very small genome sizes (Giovannoni et al. 2005; Ting et al. 2009), whereas soil microbes are often among those with the largest genomes. Our results are also in line with observations indicating different average genome sizes in distinct environments (Raes et al. 2007; Angly et al. 2009). Regarding the functional similarity of genomes, we likewise observe that it is much stronger than what would be expected based on relatedness alone (Fig. 4G). Here again, lineage-environment associations appear to be stable enough to allow selection for similar functional repertoires even in unrelated lineages.

However, apart from a phylogenetic signal, functional similarities can also arise due to similarities in genome size (van Nimwegen 2003; Konstantinidis and Tiedje 2004; Ranea et al. 2004). When correcting for the dependency between genome size

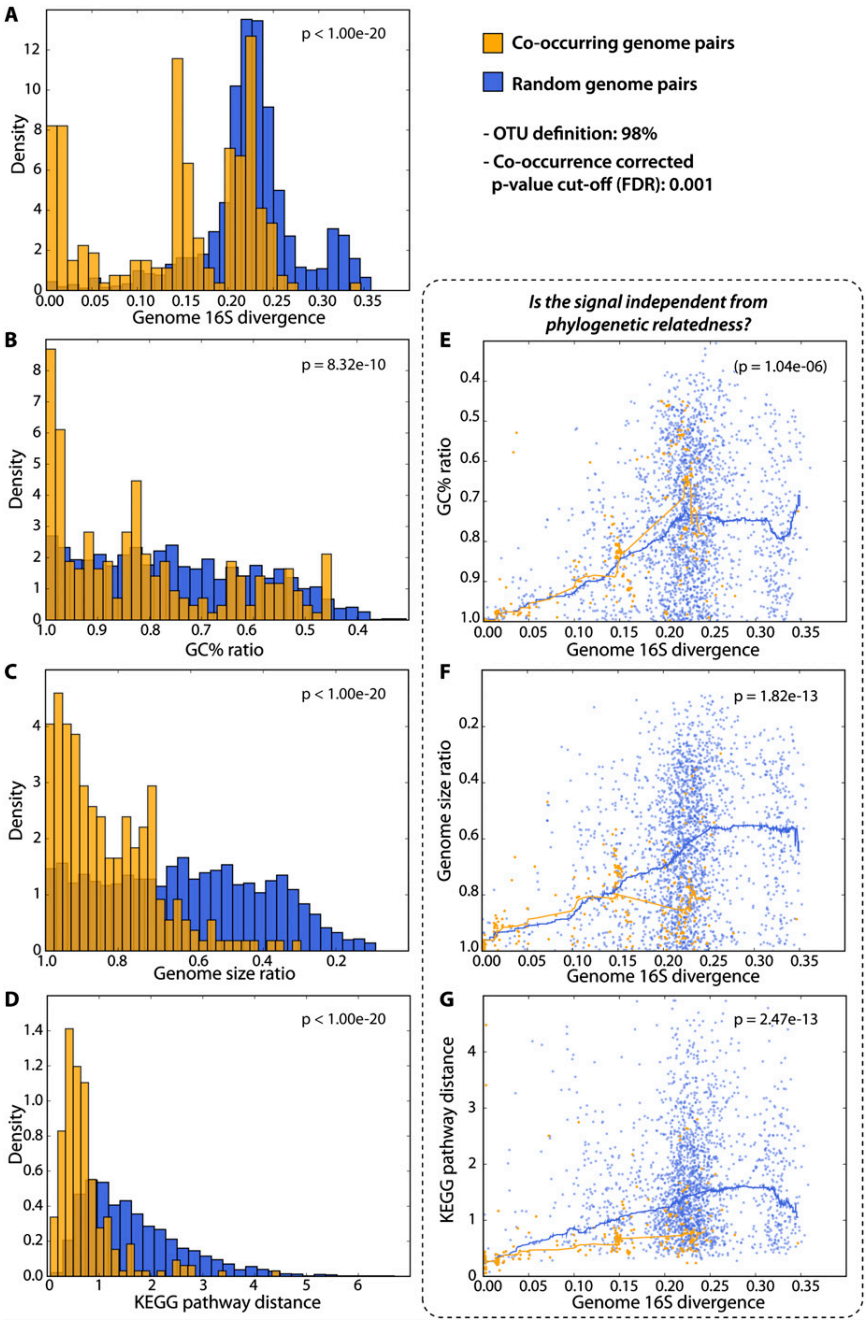


Figure 4. Coexisting lineages display similarities in genomic features. Here, we exclusively focus on co-occurring lineages for which completely sequenced genomes could be mapped to both partners (this genome mapping is globally visualized in Supplemental Fig. S6). Properties of such co-occurring genomes are compared, and contrasted against randomly paired genomes. (*A*) The distribution of 16S sequence divergence scores; shifted to the *left* for co-occurring genome pairs (i.e. they tend to be related phylogenetically). In panels *E*, *F*, and *G*, we test for independence between phylogenetic relatedness, and observations as shown in panels *B*, *C*, and *D*, respectively. Here, each dot denotes a pairwise genome comparison, and lines correspond to running medians.

and functional content, we again find that co-occurring genomes of identical size are much more similar in functional terms than expected (Fig. 5). In Figure 5, we not only plotted genome size and functional similarity, but also phylogenetic relatedness (by means of a color code). This reveals the expected, combined trends: Environmentally associated lineages that tend to be most similar in functional terms also tend to be those that are both, phylogenetically the most related and also the most similar in terms of genome size. Outliers from these trends should reveal interesting exceptions, inviting speculations on distinct ecological scenarios. We highlight a few of such extremes in Figure 5. The first example represents an outlier case because the two lineages are very closely

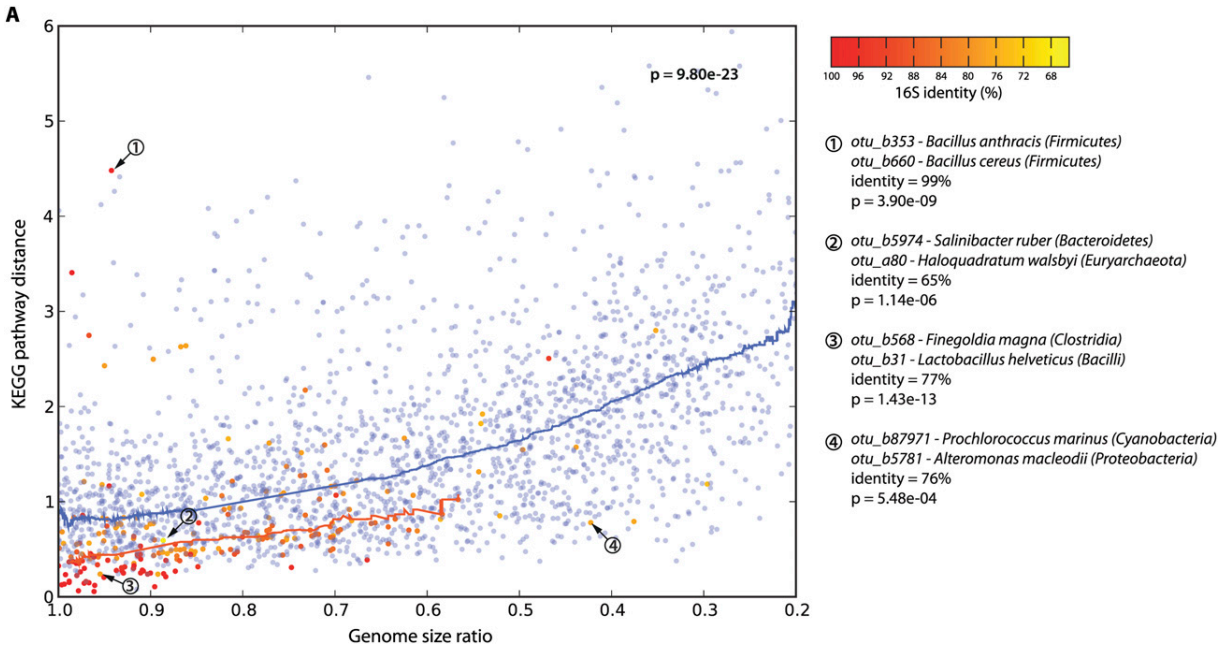


Figure 5. Functional similarities among co-occurring genomes. Each dot denotes a pair of genomes, which are either co-occurring in the environment (red to orange dots) or randomly paired (blue dots). The plot shows differences in functional genome content (*y*-axis), and in genome size (*x*-axis). Lines denote running medians. Note that, in general, the more divergent two genomes are in size, the more they are functionally distinct (blue line). In co-occurring genomes, this trend is strongly shifted toward similar functions, at all levels of phylogenetic relatedness (color-coded from red to orange). Examples of genome pairs that are discussed in the text are indicated.

related phylogenetically (both are *Bacilli*), yet they currently hold the record in terms of functional divergence. Note that the various species of *Bacilli* are very closely related phylogenetically and cannot easily be distinguished based on 16S alone (Vilas-Boas et al. 2007; Kolsto et al. 2009) (our algorithm thus assigned the two genomes arbitrarily among the two lineages, well within 98% sequence identity). Nevertheless, this reflects on the known, large phenotypic and genomic diversification within the so-called *Bacillus cereus* "group" (Vilas-Boas et al. 2007; Kolsto et al. 2009) (e.g., *Bacillus anthracis* is usually nonmotile and produces a capsule and toxins, whereas *B. cereus* tends to be motile and to make no capsule). Importantly, our data show that such a high level of phenotypic and genomic plasticity among co-occurring lineages is exceptional, especially when they are so closely related phylogenetically. Perhaps the unusual life cycle of *Bacilli* (involving a resilient endospore stage) is conducive to unusually large changes in lifestyle and phenotypes, over short time periods. In contrast, the second example describes two lineages that are very distant phylogenetically (one is an *Archaeon* and the other a *Bacterium*), and yet they co-occur quite specifically. In our data, these two (*Salinibacter* and *Haloquadratum*) are outliers because, despite their distance, they have very similar genome sizes and very similar functional pathway coverage, marking the current record at such a large phylogenetic distance. Perhaps, both lineages have independently entered the same niche (i.e., warm, fully oxygenated brines) (see also Kunin et al. 2008b), and have thus converged toward coarsely similar overall genomic features. The next example again presents two fairly unrelated lineages (related only at the phylum level), which are even more closely matched in terms of genome size and functional genome composition. They occur together very specifically, on acidic human skin, dust, and in filtered air (seven observations by five distinct laboratories; $P = 6 \times 10^{-13}$; note that dust and air samples are related to skin since they may contain small skin-derived particles). Despite apparently occupying the same niche, it is notable that only one of them (*Finegoldia*) has a tendency for opportunistic pathogenesis (Goto et al. 2008). The last example concerns two lineages that occur together in the open ocean (*Prochlorococcus* and *Alteromonas*). They are unrelated phylogenetically, and they were chosen as outliers here because their genomes are unusually distinct in size (*Alteromonas* is more than twice as large as *Prochlorococcus*) (Rocap et al. 2003) (note that our analysis has insufficient resolution to specify the exact "ecotype" for either lineage). Since both were sampled in the open water, it is difficult to envisage any mechanism of their association, but this particular pair has been noted before—*Alteromonas* has been enriched as a co-contaminant in *Prochlorococcus* cultures, facilitating the growth of the latter (Morris et al. 2008) (perhaps by alleviating oxidative stress). In the ocean, the association is probably rather unspecific, although it is conceivable that *Alteromonas*, as an opportunistically growing heterotroph, may profit from biomass accumulated by the primary producer *Prochlorococcus*. These lifestyles are quite distinct and might explain the unusually large differences in genome size.

Overall, however, we find that co-occurring genomes tend to closely match each other's genome sizes and broad functional composition. These results seem to be compatible with a picture of competition (Hibbing et al. 2010), rather than cooperation, among most of the distinct microbial lineages found at any given site: If the majority of lineages were to routinely cooperate by specialization and division of tasks, this would presumably result in genomic features that might become more distinct from each other over time. Of course, the broad view that we take here could easily make us miss cooperation among a subset of lineages, such as syntrophy and other mutual benefits from the juxtaposition of

distinct molecular capabilities. But such interactions are perhaps anyway more fleeting encounters rather than stable mutualisms. Indeed, long-term obligatory mutualism usually requires stable and specific physical contact between the organisms in question (Boucher 1985), a requirement that makes it perhaps less feasible for microbes that are generally dispersed easily (except, of course, when vertically inherited together within a eukaryote) (Vautrin and Vavre 2009).

In the future, statistical approaches like ours stand to benefit greatly from the projected further increases in both microbial genome sequencing (Ahmed 2009; Chain et al. 2009; The NIH HMP Working Group 2009) and 16S rRNA sampling (Tringe and Hugenholtz 2008; Costello et al. 2009). Both types of data will prove particularly valuable when augmented with standardized information about the environments sampled, for example, by following the recommendations of the MIENS standard (http://gensc.org/gc_wiki/index.php/MIENS). Novel and specific microbial assemblages can already be identified using the current data (see Fig. 3; Supplemental material), and more such discoveries can be expected with higher data coverage. Note that our approach does not require prior information about environmental ontologies or hierarchies of sampling sites; instead, groups of biologically related sampling sites are defined by the data themselves (Fig. 2; Supplemental Fig. S11). In general, approaches that integrate sequence data from both strain sequencing and from environmental marker gene sequencing hold great potential, since they connect the molecular information contained in the (pan-)genome of each lineage to the quantitative occurrence pattern of that lineage around the globe.

Methods

Definition of taxonomic units and sampling events

All 298,591 available 16S rRNA sequence records were downloaded from the Greengenes database (Desantis et al. 2006) on January 2009. At Greengenes, these sequences had already been cleaned of potential chimera by the program Bellerophon (Huber et al. 2004). We filtered sequences according to their lengths (≥900 nt for Archaea and ≥1200 for Bacteria) and additionally flagged sequences predicted to be chimeric by the program ChimeraSlayer (http://microbiomeutil.sourceforge.net/). We also removed from the analysis all sequences lacking annotations in any of the fields "author," "title," or "isolation_source." This was done in order to be able to define a sampling event for each record. In our study, a "sampling event" is defined as the unique concatenation of these three annotation fields (author + title + isolation_source).

Archaeal and bacterial sequences were aligned separately, using the secondary-structure aware aligner "Infernal" (Nawrocki et al. 2009), together with the corresponding 16S rRNA covariance models of the RDP database (Cole et al. 2009). Before defining OTUs, we removed sequences for which the alignment had not been successful (i.e., Infernal bit-score < 0). OTUs were built for both Archaea and Bacteria by hierarchical clustering (complete linkage), at various distances (from 0.2 to 0.01), using the clustering tool of the RDP pyrosequencing pipeline (Cole et al. 2009; http://pyro.cme.msu.edu/). Because not all 16S sequences reported in databases are necessarily genuine environmental sequences (Tanner et al. 1998; Barton et al. 2006), we assembled a database of potential laboratory contaminants, containing 38 distinct sequences (Tanner et al. 1998; Barton et al. 2006). Homology searches revealed that between 47 and 309 of our OTUs contained such sequences (matching at 97% sequence identity or better). However, these OTUs are rarely involved in significant co-occurrences; for example, in Figure 2 only 17 of the OTUs shown contain potential contaminants, and these are scattered over various smaller clusters (they are flagged in Fig. 2 and in the detailed Supplemental material).

In order to compute sequence divergence values for pairs of OTUs, we first selected a single sequence to represent each OTU. (We chose the sequence that had the minimum sum of squares of distances to all other sequences within that cluster; note that this does not favor short sequences since the distances we used are length-normalized.) We then aligned these representative sequences pairwise (using the program "water" from the EMBOSS package) (Rice et al. 2000) and determined their sequence identity.

Classification of taxonomic units

In order to assign taxonomic classifications to entire OTUs, we first assessed the pre-annotated taxonomies of all individual 16S rRNA sequences in Greengenes (*sensu* RDP taxonomy). Where these were still annotated as "unclassified," we re-ran the taxonomy classification using the RDP classifier (Cole et al. 2009). Taxonomy predictions reported there were considered reliable, if supported by a minimum bootstrap value of 80%. To assign taxonomy classifications to OTUs, we then used a simple majority vote: If more than half of the sequences present within a cluster agreed upon a classification, the OTU was annotated as belonging to this taxon. In case of conflicts, we assigned consensus classifications at increasingly higher levels of taxonomy until the majority vote condition was again met.

Co-occurrence analysis

In order to reduce the search space for co-occurrence testing (which encompasses potentially more than 2 billion pairs, for example, in the case of OTUs defined at 99% sequence identity), we limited our analysis to OTUs occurring in at least three distinct sampling sites. Conversely, we only considered sampling events encompassing at least two distinct OTUs. For these "filtered" OTUs and samplings (see also Table 1), we tested the co-occurrence significance for all possible pairs using the Fisher's exact test. For each test, the four cells in the contingency table denoted the number of samples containing both OTUs, one of the two OTUs only, or none of the two, respectively. Subsequently, we adjusted all *P*-values for multiple testing using the Benjamini and Hochberg FDR controlling procedure (Benjamini and Hochberg 1995), as implemented in the "*multtest*" library of the statistical software package R (http://www.r-project.org). We also verified our FDR empirically, by re-computing the associations using randomized input data. For this, we randomly reassigned the various OTUs to the various sampling events, under the constraint that each OTU kept the overall number of samples it mapped to, and each sample kept the overall number of OTUs. This maintained the size distributions of both, samples and OTUs (results are provided in Table 1). To compute the necessary large number of tests in a reasonable time, we used a C-implementation of the test in the Apophenia library for scientific computing (http://apophenia.sourceforge.net), using the python SWIG interface as a wrapper. For selected examples of co-occurring lineages discussed in the text (Figs. 1, 3; Supplemental Fig. S1), we also computed the odds-ratio ("or," a statistical measure of effect size), in order to assess the strength of the reported associations. Note that our input data, and thus also our predicted associations, likely suffer from undersampling and probably also from systematic biases in the sampling. Both effects are difficult to quantify, but are likely present due to variable choices of PCR primers (information about primers

is often not available in the sequence records), and also due to experimental biases in DNA extraction protocols. However, while such biases can likely suppress the detection of certain lineages, it is less likely that they generate false-positive associations at the level of specificity that we observe here (see Supplemental Fig. S3, and see also the randomizations described above). We also noted that, overall, larger samples contribute more co-occurrence associations than smaller samples, as expected. We quantified this in two ways: by stratifying the input data by sample size, and by randomly down-sampling the larger environmental samples (these often focus on the mammalian gut). The results of both tests are summarized in Supplemental Figure S7; reassuringly, we observe that entirely removing gut-related samples through keyword searches, while lowering the number of association clusters, still supports the quantitative conclusion that we report in Figures 4 and 5 (see Supplemental Fig. S10).

Network inference

Based on the co-occurrence analysis results, we constructed networks of coexisting microbes for different levels of OTU definitions. For this, the FDR cutoff for each individual edge in the network was 0.001. In order to obtain a simplified view on the results and to identify cohesive modules of coexisting microbial lineages, we clustered our networks using the Markov cluster algorithm (MCL algorithm; http://micans.org/mcl) (Enright et al. 2002). This clustering was performed using as the similarity metric (i.e., edge weights) the normalized co-occurrence similarity between OTUs, defined here as the Jaccard similarity coefficient (i.e., cooc_count/[(otu1_count + otu2_count) − cooc_count]). We set MCL's "inflation" parameter to 2.0 when running the algorithm. All network images were generated using custom Python scripting and the Python module "NetworkX" (http://networkx.lanl.gov), which provides an interface to the "Graphviz" graph visualization software (http://www.graphviz.org).

Cluster annotation

To annotate clusters in the co-occurrence network with environmental information, we relied on the controlled vocabulary maintained by the Environment Ontology project (EnvO, version 1.51; http://environmentontology.org). In a first step, we assigned EnvO keywords to each OTU in the network; to do so, we scanned all words in the "isolation_source" field from each OTU and assigned ontology terms to that OTU based on exact matches. For many of its terms, EnvO also provides "synonyms"; for cases in which a term could not be matched directly, we also allowed matches via these synonyms, but only for synonyms of the categories "EXACT" or "NARROW" (omitting the categories "RELATED" and "BROAD"). The Fisher's exact test then allowed us to assign significantly over-represented keywords (FDR = 0.01; *P*-value adjusted for multiple testing using the Benjamini and Hochberg procedure) for each given cluster or subnetwork, compared to the background frequency of these terms in the entire network.

Comparative genomics

First, we mapped available, completely sequenced genomes to our OTUs, for various levels of OTU definitions. For this, we extracted the 16S rRNA genes predicted for 881 complete genomes contained in the RefSeq database (RefSeq 35, 05-13-2009), requiring a minimum length of 700 bp. We then compared these sequences against representative 16S sequences from each OTU, using BLAST with the following parameters: "-a 2 -m 8 -p blastn -v 1000 -b 1000 -r 2 -q -3 -G 5 -E 2 -e 0.01." For genomes that are annotated with more than one predicted 16S rRNA gene, we retained the longest copy. For the mapping, we then ranked all sequence matches by bit-score (best score first) and, parsing through the list, assigned each genome to the best-matching OTU (skipping those that were already previously assigned to another genome). In addition, we required that the alignment length for the BLAST hit was at least 800 bp and that the sequence identity of the match was 97% or greater.

We then analyzed co-occurring OTUs by comparing their mapped genomes, using several characteristics: genome size, GC content, and relative coverage of KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways (Kanehisa et al. 2008). To compute genome size ratios, we used the total DNA length of the non-redundant chromosomes and plasmids, expressed in nucleotides; to compute GC content ratios, we used the predetermined values for the complete genomes available at http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi. In order to compare genomes in terms of their encoded functions, we assessed the relative coverage of pathways as annotated at KEGG, using the KEGG API (http://www.genome.jp/kegg/soap). We computed normalized vectors describing the relative pathway coverage among all annotated genes of a given genome and then compared these vectors by computing their Euclidean distance. In order to exclude potential artifacts arising from occasional annotation errors in KEGG, we repeated this analysis with two additional, independent systems of functional genome annotation, retrieving essentially the same results (Supplemental Fig. S4).

To test for the statistical independence of our observations made for a given distance measure, against another measure (usually against phylogenetic distance) (Fig. 4E–G), we first learned the dependency between the two measures based on randomly selected pairs of genomes (blue dots). This dependency was then described using a running median (blue lines in Fig. 4). Next, we assessed the data of interest (i.e., pairs of co-occurring genomes) by computing for each data point its vertical distance to the (blue) running median, divided by that median itself. This measure has been termed "relative distance to median" ("dm"; see, for example, Newman et al. 2006); it permits us to compare data at a given, fixed setting of a second, potentially confounding variable. From this, we generated a distribution of normalized distance values, which we compared to the corresponding random background distributions, using the non-parametric Kolmogorov-Smirnov test.

Data availability

Raw input data, as well as all computed results of this study (including sequence data, operational taxonomic units, co-occurrence statistics, network clustering, and genome mapping) are available online at http://mblnx-kallisto.uzh.ch:8888/microbial_coexistence/. In addition, a zoomable and clickable version of the network in Figure 2C is available as Supplemental Figure S12, which can be downloaded from the Supplemental materials.

Acknowledgments

This work was funded by the Swiss National Science Foundation and by the University of Zurich through its Research Priority Program in Systems Biology and Functional Genomics. We thank Phil Hugenholtz and Todd DeSantis for help with the Greengenes database, and Wolf-Dietrich Hardt for insightful comments and criticism.

References

Ahmed N. 2009. A flood of microbial genomes—do we need more? *PLoS One* **4:** e5831. doi: 10.1371/journal.pone.0005831.

Alonso C, Warnecke F, Amann R, Pernthaler J. 2007. High local and global diversity of Flavobacteria in marine plankton. *Environ Microbiol* **9:** 1253–1266.

Angly FE, Willner D, Prieto-Davo A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, et al. 2009. The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol* **5:** e1000593. doi: 10.1371/journal.pcbi.1000593.

Baati H, Guermazi S, Amdouni R, Gharsallah N, Sghir A, Ammar E. 2008. Prokaryotic diversity of a Tunisian multipond solar saltern. *Extremophiles* **12:** 505–518.

Barabasi AL, Oltvai ZN. 2004. Network biology: Understanding the cell's functional organization. *Nat Rev Genet* **5:** 101–113.

Barton HA, Taylor NM, Lubbers BR, Pemberton AC. 2006. DNA extraction from low-biomass carbonate rock: An improved method with reduced contamination and the low-biomass contaminant database. *J Microbiol Methods* **66:** 21–31.

Bell T, Ager D, Song JI, Newman JA, Thompson IP, Lilley AK, van der Gast CJ. 2005. Larger islands house more bacterial taxa. *Science* **308:** 1884.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol* **57:** 289–300.

Boetius A, Ravenschlag K, Schubert CJ, Rickert D, Widdel F, Gieseke A, Amann R, Jorgensen BB, Witte U, Pfannkuche O. 2000. A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* **407:** 623–626.

Boucher DH. 1985. *The biology of mutualism: Ecology and evolution*. Oxford University Press, New York.

Brauman A, Kane MD, Labat M, Breznak JA. 1992. Genesis of acetate and methane by gut bacteria of nutritionally diverse termites. *Science* **257:** 1384–1387.

Caldwell SL, Laidler JR, Brewer EA, Eberly JO, Sandborgh SC, Colwell FS. 2008. Anaerobic oxidation of methane: Mechanisms, bioenergetics, and the ecology of associated microorganisms. *Environ Sci Technol* **42:** 6791–6799.

Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, et al. 2009. Genomics. Genome project standards in a new era of sequencing. *Science* **326:** 236–237.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, et al. 2009. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37:** D141–D145.

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. *Science* **326:** 1694–1697.

de Bary A. 1879. *Die Erscheinung der Symbiose*. Verlag Karl J. Trubner, Strassbourg.

Desantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72:** 5069–5072.

de Wit R, Bouvier T. 2006. "Everything is everywhere, but, the environment selects"; what did Baas Becking and Beijerinck really say? *Environ Microbiol* **8:** 755–758.

Dubilier N, Mulders C, Ferdelman T, de Beer D, Pernthaler A, Klein M, Wagner M, Erseus C, Thiermann F, Krieger J, et al. 2001. Endosymbiotic sulphate-reducing and sulphide-oxidizing bacteria in an oligochaete worm. *Nature* **411:** 298–302.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30:** 1575–1584.

Ferriere R, Gauduchon M, Bronstein JL. 2007. Evolution and persistence of obligate mutualists and exploiters: Competition for partners and evolutionary immunization. *Ecol Lett* **10:** 115–126.

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, et al. 2008a. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* **26:** 541–547.

Field D, Garrity GM, Sansone SA, Sterk P, Gray T, Kyrpides N, Hirschman L, Glockner FO, Kottmann R, Angiuoli S, et al. 2008b. Meeting report: The fifth Genomic Standards Consortium (GSC) workshop. *OMICS* **12:** 109–113.

Finlay BJ. 2002. Global dispersal of free-living microbial eukaryote species. *Science* **296:** 1061–1063.

Forterre P, Gribaldo S, Brochier-Armanet C. 2009. Happy together: Genomic insights into the unique *Nanoarchaeum/Ignicoccus* association. *J Biol* **8:** 7. doi: 10.1186/jbiol110.

Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, et al. 1980. The phylogeny of prokaryotes. *Science* **209:** 457–463.

Freilich S, Kreimer A, Meilijson I, Gophna U, Sharan R, Ruppin E. 2010. The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Res* doi: 10.1093/nar/gkq118.

Fuhrman JA. 2009. Microbial community structure and its functional implications. *Nature* **459:** 193–199.

Fuhrman JA, Steele JA. 2008. Community structure of marine bacterioplankton: Patterns, networks, and relationships to function. *Aquat Microb Ecol* **53:** 69–81.

Garcia-Pichel F, Nubel U, Muyzer G. 1998. The phylogeny of unicellular, extremely halotolerant cyanobacteria. *Arch Microbiol* **169:** 469–482.

Ghosh A, Dey N, Bera A, Tiwari A, Sathyaniranjan K, Chakrabarti K, Chattopadhyay D. 2010. Culture independent molecular analysis of bacterial communities in the mangrove sediment of Sundarban, India. *Saline Systems* **6:** 1. doi: 10.1186/1746-1448-6-1.

Gibbs KA, Urbanowski ML, Greenberg EP. 2008. Genetic determinants of self identity and social recognition in bacteria. *Science* **321:** 256–259.

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. 1990. Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345:** 60–63.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, et al. 2005. Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309:** 1242–1245.

Goto T, Yamashita A, Hirakawa H, Matsutani M, Todo K, Ohshima K, Toh H, Miyamoto K, Kuhara S, Hattori M, et al. 2008. Complete genome sequence of *Finegoldia magna*, an anaerobic opportunistic pathogen. *DNA Res* **15:** 39–47.

Hamady M, Knight R. 2009. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res* **19:** 1141–1152.

Hansen S, Rainey P, Haagensen J, Molin S. 2007. Evolution of species interactions in a biofilm community. *Nature* **445:** 533–536.

Harcombe W. 2010. Novel cooperation experimentally evolved between species. *Evolution*. doi: 10.1111/j.1558-5646.2010.00959.x.

Harrison BK, Zhang H, Berelson W, Orphan VJ. 2009. Variations in archaeal and bacterial diversity associated with the sulfate-methane transition zone in continental margin sediments (Santa Barbara Basin, California). *Appl Environ Microbiol* **75:** 1487–1499.

Harvey PH, Colwell RK, Silvertown JW, May RM. 1983. Null models in ecology. *Annu Rev Ecol Syst* **14:** 189–211.

Heijs SK, Damste JS, Forney LJ. 2005. Characterization of a deep-sea microbial mat from an active cold seep at the Milano mud volcano in the Eastern Mediterranean Sea. *FEMS Microbiol Ecol* **54:** 47–56.

Hibbing ME, Fuqua C, Parsek MR, Peterson SB. 2010. Bacterial competition: Surviving and thriving in the microbial jungle. *Nat Rev Microbiol* **8:** 15–25.

Horner-Devine MC, Silver JM, Leibold MA, Bohannan BJ, Colwell RK, Fuhrman JA, Green JL, Kuske CR, Martiny JB, Muyzer G, et al. 2007. A comparison of taxon co-occurrence patterns for macro- and microorganisms. *Ecology* **88:** 1345–1353.

Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO. 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417:** 63–67.

Huber T, Faulkner G, Hugenholtz P. 2004. Bellerophon: A program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20:** 2317–2319.

Hubert C, Loy A, Nickel M, Arnosti C, Baranyi C, Bruchert V, Ferdelman T, Finster K, Christensen FM, Rosa de Rezende J, et al. 2009. A constant flux of diverse thermophilic bacteria into the cold Arctic seabed. *Science* **325:** 1541–1544.

Inagaki F, Nunoura T, Nakagawa S, Teske A, Lever M, Lauer A, Suzuki M, Takai K, Delwiche M, Colwell FS, et al. 2006. Biogeographical distribution and diversity of microbes in methane hydrate-bearing deep marine sediments on the Pacific Ocean Margin. *Proc Natl Acad Sci* **103:** 2815–2820.

Isenbarger TA, Finney M, Rios-Velazquez C, Handelsman J, Ruvkun G. 2008. Miniprimer PCR, a new lens for viewing the microbial world. *Appl Environ Microbiol* **74:** 840–849.

Johnstone RA, Bshary R. 2008. Mutualism, market effects and partner control. *J Evol Biol* **21:** 879–888.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, et al. 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36:** D480–D484.

Knittel K, Boetius A. 2009. Anaerobic oxidation of methane: Progress with an unknown process. *Annu Rev Microbiol* **63:** 311–334.

Kolsto AB, Tourasse NJ, Okstad OA. 2009. What sets *Bacillus anthracis* apart from other *Bacillus* species? *Annu Rev Microbiol* **63:** 451–476.

Kolter R, Greenberg EP. 2006. Microbial sciences: The superficial life of microbes. *Nature* **441:** 300–302.

Konstantinidis KT, Tiedje JM. 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci* **101:** 3160–3165.

Kottmann R, Kostadinov I, Duhaime MB, Buttigieg PL, Yilmaz P, Hankeln W, Waldmann J, Glockner FO. 2010. Megx.net: Integrated database resource for marine ecological genomics. *Nucleic Acids Res* **38:** D391–D395.

Kunin V, He S, Warnecke F, Peterson SB, Garcia Martin H, Haynes M, Ivanova N, Blackall LL, Breitbart M, Rohwer F, et al. 2008a. A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Res* **18:** 293–297.
Kunin V, Raes J, Harris JK, Spear JR, Walker JJ, Ivanova N, von Mering C, Bebout BM, Pace NR, Bork P, et al. 2008b. Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Mol Syst Biol* **4:** 198. doi: 10.1038/msb.2008.35.
Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci* **82:** 6955–6959.
Langenheder S, Prosser JI. 2008. Resource availability influences the diversity of a functional group of heterotrophic soil bacteria. *Environ Microbiol* **10:** 2245–2256.
Langenheder S, Lindstrom ES, Tranvik LJ. 2006. Structure and function of bacterial communities emerging from different sources under identical conditions. *Appl Environ Microbiol* **72:** 212–220.
Ley RE, Harris JK, Wilcox J, Spear JR, Miller SR, Bebout BM, Maresca JA, Bryant DA, Sogin ML, Pace NR. 2006. Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat. *Appl Environ Microbiol* **72:** 3685–3695.
Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. 2008. Worlds within worlds: Evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* **6:** 776–788.
Li T, Wang P. 2008. [Bacterial and archaeal diversity in surface sediment from the south slope of the South China Sea]. *Wei Sheng Wu Xue Bao* **48:** 323–329.
Li T, Wang P, Wang PX. 2008. Microbial diversity in surface sediments of the Xisha Trough, the South China Sea. *Acta Ecologica Sinica* **28:** 1166–1173.
Lloyd KG, Lapham L, Teske A. 2006. An anaerobic methane-oxidizing community of ANME-1b archaea in hypersaline Gulf of Mexico sediments. *Appl Environ Microbiol* **72:** 7218–7230.
Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IM, Grechkin Y, Dubchak I, Anderson I, et al. 2008. IMG/M: A data management and analysis system for metagenomes. *Nucleic Acids Res* **36:** D534–D538.
Martiny JB, Bohannan BJ, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, et al. 2006. Microbial biogeography: Putting microorganisms on the map. *Nat Rev Microbiol* **4:** 102–112.
McHardy AC, Rigoutsos I. 2007. What's in the mix: Phylogenetic classification of metagenome sequence samples. *Curr Opin Microbiol* **10:** 499–503.
Min KR, Rickard AH. 2009. Coaggregation by the freshwater bacterium *Sphingomonas natatoria* alters dual-species biofilm formation. *Appl Environ Microbiol* **75:** 3987–3997.
Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet* **17:** 589–596.
Morris JJ, Kirkegaard R, Szul MJ, Johnson ZI, Zinser ER. 2008. Facilitation of robust growth of *Prochlorococcus* colonies and dilute liquid cultures by "helper" heterotrophic bacteria. *Appl Environ Microbiol* **74:** 4530–4534.
Mrazek J. 2009. Phylogenetic signals in DNA composition: Limitations and prospects. *Mol Biol Evol* **26:** 1163–1169.
Muyzer G, Teske A, Wirsen CO, Jannasch HW. 1995. Phylogenetic relationships of *Thiomicrospira* species and their identification in deep-sea hydrothermal vent samples by denaturing gradient gel electrophoresis of 16S rDNA fragments. *Arch Microbiol* **164:** 165–172.
Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: Inference of RNA alignments. *Bioinformatics* **25:** 1335–1337.
Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS. 2006. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441:** 840–846.
Newton RJ, Jones SE, Helmus MR, McMahon KD. 2007. Phylogenetic ecology of the freshwater *Actinobacteria* acI lineage. *Appl Environ Microbiol* **73:** 7169–7176.
Ng WL, Bassler BL. 2009. Bacterial quorum-sensing network architectures. *Annu Rev Genet* **43:** 197–222.
The NIH HMP Working Group. 2009. The NIH Human Microbiome Project. *Genome Res* **19:** 2317–2323.
Nilsson AI, Koskiniemi S, Eriksson S, Kugelberg E, Hinton JC, Andersson DI. 2005. Bacterial genome size reduction by experimental evolution. *Proc Natl Acad Sci* **102:** 12112–12116.
Orphan VJ. 2009. Methods for unveiling cryptic microbial partnerships in nature. *Curr Opin Microbiol* **12:** 231–237.
Overmann J, Schubert K. 2002. Phototrophic consortia: Model systems for symbiotic interrelations between prokaryotes. *Arch Microbiol* **177:** 201–208.
Palmer TM, Stanton ML, Young TP. 2003. Competition and coexistence: Exploring mechanisms that restrict and maintain diversity within mutualist guilds. *Am Nat* **162:** S63–S79.
Paracer S, Ahmadjian V. 2000. *Symbiosis: An introduction to biological associations*. Oxford University Press, New York.
Parsek MR, Greenberg EP. 2005. Sociomicrobiology: The connections between quorum sensing and biofilms. *Trends Microbiol* **13:** 27–33.
Pedros-Alio C. 2006. Marine microbial diversity: Can it be determined? *Trends Microbiol* **14:** 257–263.
Piccini C, Conde D, Alonso C, Sommaruga R, Pernthaler J. 2006. Blooms of single bacterial species in a coastal lagoon of the southwestern Atlantic Ocean. *Appl Environ Microbiol* **72:** 6560–6568.
Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO. 2007. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35:** 7188–7196.
Raes J, Bork P. 2008. Molecular eco-systems biology: Towards an understanding of community function. *Nat Rev Microbiol* **6:** 693–699.
Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P. 2007. Prediction of effective genome size in metagenomic samples. *Genome Biol* **8:** R10. doi: 10.1186/gb-2007-8-1-r10.
Ranea JA, Buchan DW, Thornton JM, Orengo CA. 2004. Evolution of protein superfamilies and bacterial genome size. *J Mol Biol* **336:** 871–887.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* **16:** 276–277.
Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, et al. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424:** 1042–1047.
Ruan Q, Dutta D, Schwalbach MS, Steele JA, Fuhrman JA, Sun F. 2006. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics* **22:** 2532–2538.
Ruehland C, Blazejak A, Lott C, Loy A, Erseus C, Dubilier N. 2008. Multiple bacterial symbionts in two species of co-occurring gutless oligochaete worms from Mediterranean sea grass sediments. *Environ Microbiol* **10:** 3404–3416.
Ryan RP, Dow JM. 2008. Diffusible signals and interspecies communication in bacteria. *Microbiology* **154:** 1845–1858.
Saffo MB. 1993. Coming to terms with a field: Words and concepts in symbiosis. *Symbiosis* **14:** 17–31.
Sahl JW, Pace NR, Spear JR. 2008. Comparative molecular analysis of endoevaporitic microbial communities. *Appl Environ Microbiol* **74:** 6444–6446.
Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. 2007. CAMERA: A community resource for metagenomics. *PLoS Biol* **5:** e75. doi: 10.1371/journal.pbio.0050075.
Shank EA, Kolter R. 2009. New developments in microbial interspecies signaling. *Curr Opin Microbiol* **12:** 205–214.
Smith VH. 2007. Microbial diversity–productivity relationships in aquatic ecosystems. *FEMS Microbiol Ecol* **62:** 181–186.
Sorensen KB, Canfield DE, Teske AP, Oren A. 2005. Community composition of a hypersaline endoevaporitic microbial mat. *Appl Environ Microbiol* **71:** 7352–7365.
Staley JT, Konopka A. 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol* **39:** 321–346.
Takeuchi M, Komai T, Hanada S, Tamaki S, Tanabe S, Miyachi Y, Uchiyama M, Nakazawa T, Kimura K, Kamagata Y. 2009. Bacterial and archaeal 16S rRNA genes in Late Pleistocene to Holocene muddy sediments from the Kanto Plain of Japan. *Geomicrobiol J* **26:** 104–118.
Tanner MA, Goebel BM, Dojka MA, Pace NR. 1998. Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants. *Appl Environ Microbiol* **64:** 3110–3113.
Ting CS, Ramsey ME, Wang YL, Frost AM, Jun E, Durham T. 2009. Minimal genomes, maximal productivity: Comparative genomics of the photosystem and light-harvesting complexes in the marine cyanobacterium, *Prochlorococcus*. *Photosynth Res* **101:** 1–19.
Tokura M, Ohkuma M, Kudo T. 2000. Molecular phylogeny of methanogens associated with flagellated protists in the gut and with the gut epithelium of termites. *FEMS Microbiol Ecol* **33:** 233–240.
Tringe SG, Hugenholtz P. 2008. A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* **11:** 442–446.
Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428:** 37–43.
van Nimwegen E. 2003. Scaling laws in the functional content of genomes. *Trends Genet* **19:** 479–484.
Vautrin E, Vavre F. 2009. Interactions between vertically transmitted symbionts: Cooperation or conflict? *Trends Microbiol* **17:** 95–99.

Vilas-Boas GT, Peruca AP, Arantes OM. 2007. Biology and taxonomy of *Bacillus cereus*, *Bacillus anthracis*, and *Bacillus thuringiensis*. *Can J Microbiol* **53:** 673–687.
Visick KL, Fuqua C. 2005. Decoding microbial chatter: Cell–cell communication in bacteria. *J Bacteriol* **187:** 5507–5519.
Von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P. 2007. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315:** 1126–1130.
Wanner G, Vogl K, Overmann J. 2008. Ultrastructural characterization of the prokaryotic symbiosis in "*Chlorochromatium aggregatum*." *J Bacteriol* **190:** 3721–3730.
Ward DM, Weller R, Bateson MM. 1990. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* **345:** 63–65.
Waters C, Bassler B. 2005. Quorum sensing: Cell-to-cell communication in bacteria. *Annu Rev Cell Dev Biol* **21:** 319–346.
Woyke T, Teeling H, Ivanova N, Huntemann M, Richter M, Gloeckner F, Boffelli D, Anderson I, Barry K, Shapiro H, et al. 2006. Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443:** 950–955.
Zhang W, Ki JS, Qian PY. 2008. Microbial diversity in polluted harbor sediments I: Bacterial community assessment based on four clone libraries of 16S rDNA. *Estuarine Coastal Shelf Sci* **76:** 668–681.

Received December 22, 2009; accepted in revised form April 22, 2010.

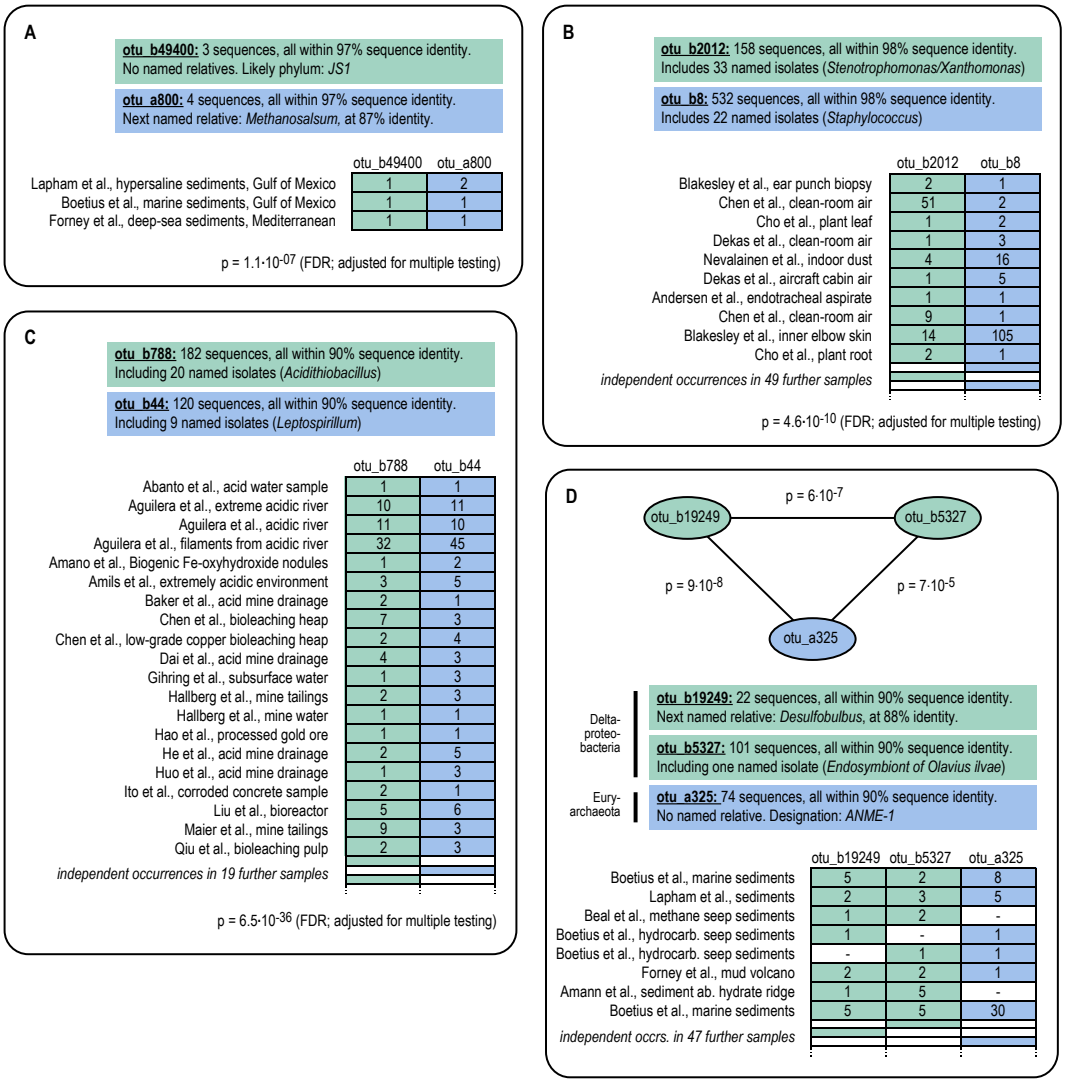


Figure S1: Additional examples of coexisting microbial lineages, as discussed in the text. The examples shown were assembled by manual searches in the co-occurrence data, at various OTU definition cutoffs (i.e., 16S rRNA sequence identity cutoffs). Note that, for technical reasons, the author name that is listed for each study is not necessarily the first author of the publication (if any), but simply the one ranking first in the corresponding database entry (usually sorted alphabetically). Adjusted p-values (FDR; p) and odds ratios (or) are reported for selected OTU pairs.

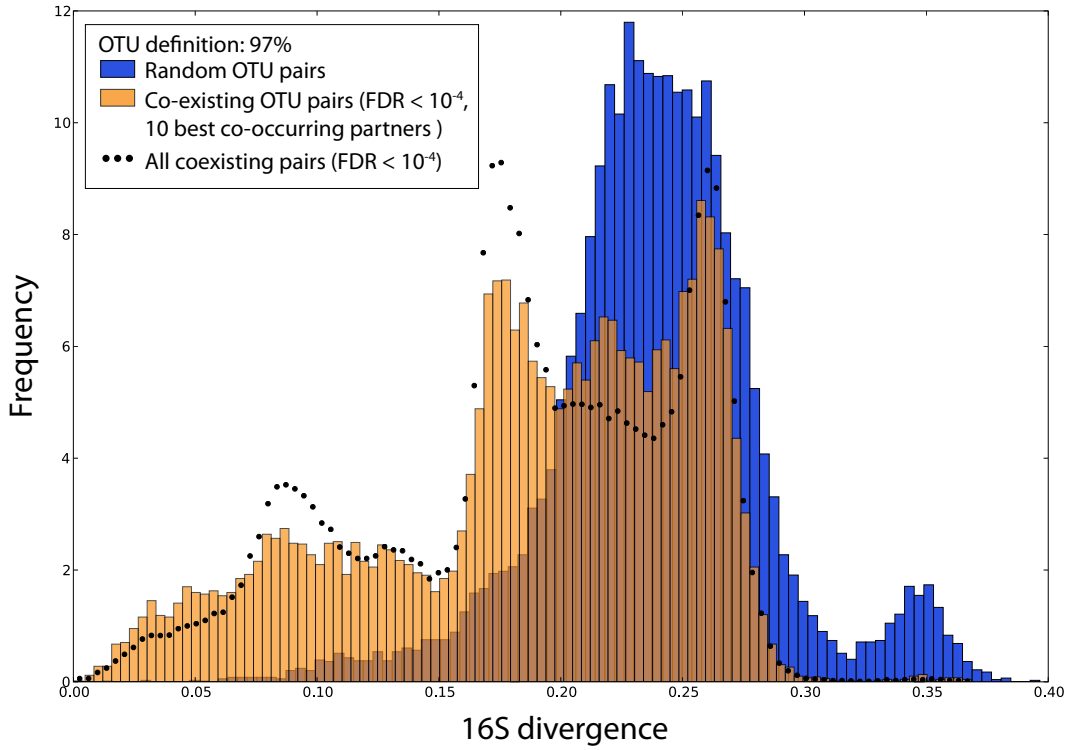


Figure S2: Phylogenetic distances between coexisting microbial lineages.
This plot shows the distribution of phylogenetic distances among co-occurring lineages, similar to what is shown in the corresponding plot in Figure 4A. However, in contrast to Figure 4A, the distribution is not limited to lineages having completely sequenced genomes, but is instead based on a comparison of all lineages using 16S sequence identity. The background distribution (randomly chosen pairs of lineages) is shown in blue, the distribution of co-occurring lineages is shown as a dotted line. Since the latter distribution is to some extent skewed by a few large clusters of co-occurring lineages, the distribution in orange is shown in addition – here, only the ten most-significant co-occurrence partners of each lineage are included.

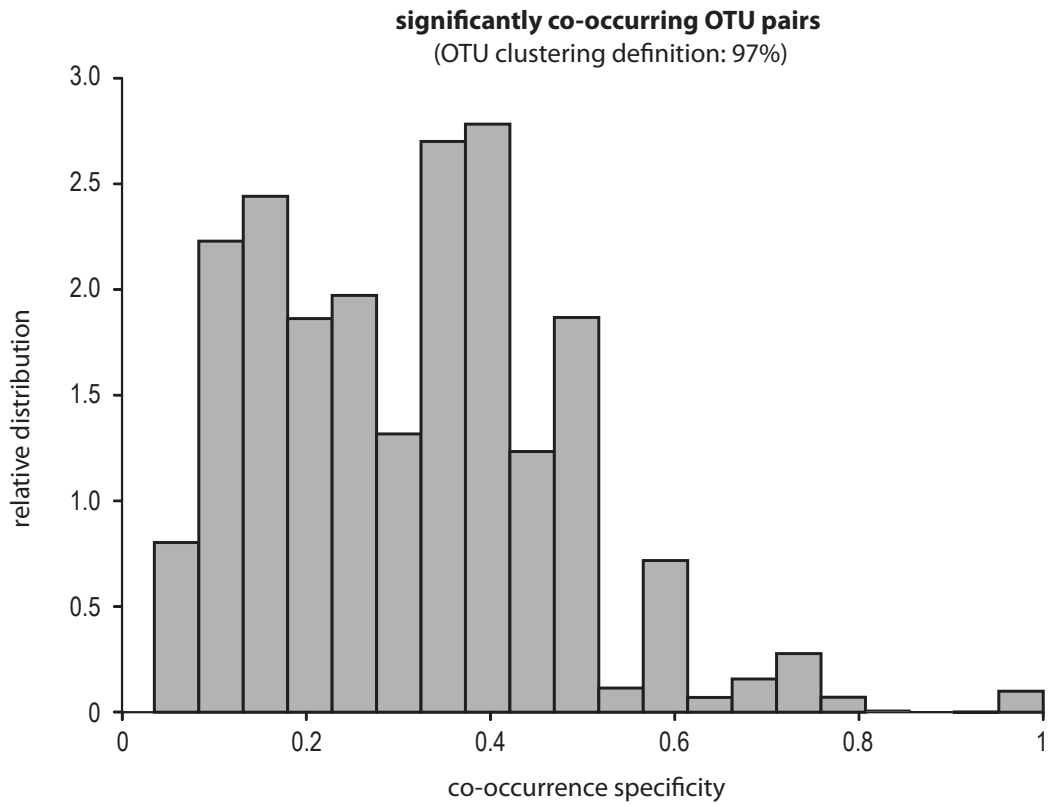


Figure S3: Distribution of association 'specificity' scores.
A co-occurrence interaction is defined to have a 'specificity' of 1.0 when the lineages in question never occur separately. The graph shows the distribution of specificity scores among all significant co-occurrences, based on OTUs defined with a clustering cutoff of 97% identity.

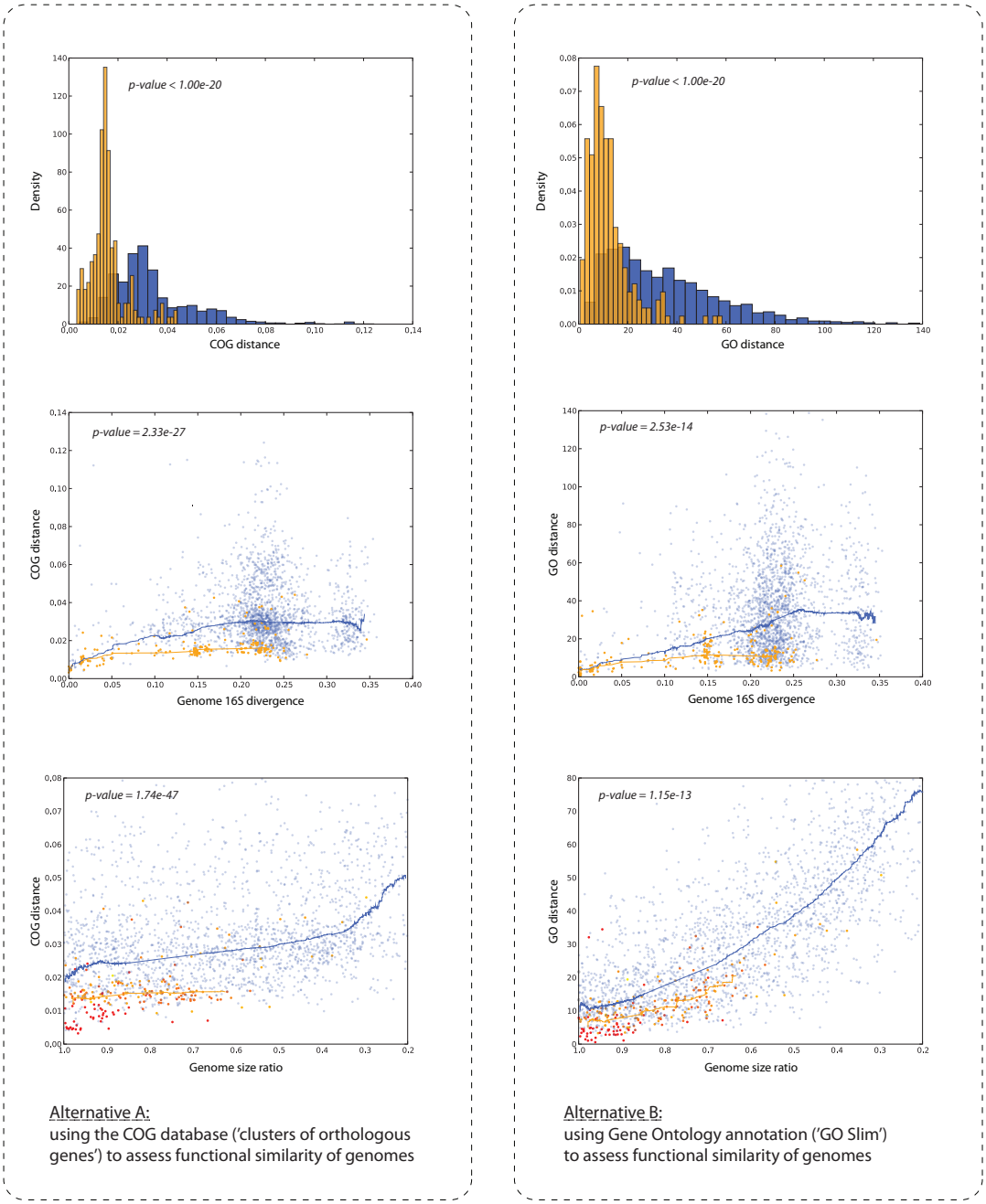


Figure S4: Functional similarity of associated lineages according to Gene Ontology or COG.

In Figures 4 and 5 of the main manuscript, functional annotations from the KEGG database were used to assess functional similarity among co-existing genomes. To exclude potential artifacts stemming from occasional annotation errors in KEGG, we repeated the analysis using two different functional annotation systems: COG ('Clusters of Orthologous Groups'), and GO ('Gene Ontology'). The figure reproduces the panels 4D, 4G and 5 for both systems, with essentially identical results (the COG-based graphs actually show the highest separation between co-existing and control genomes).

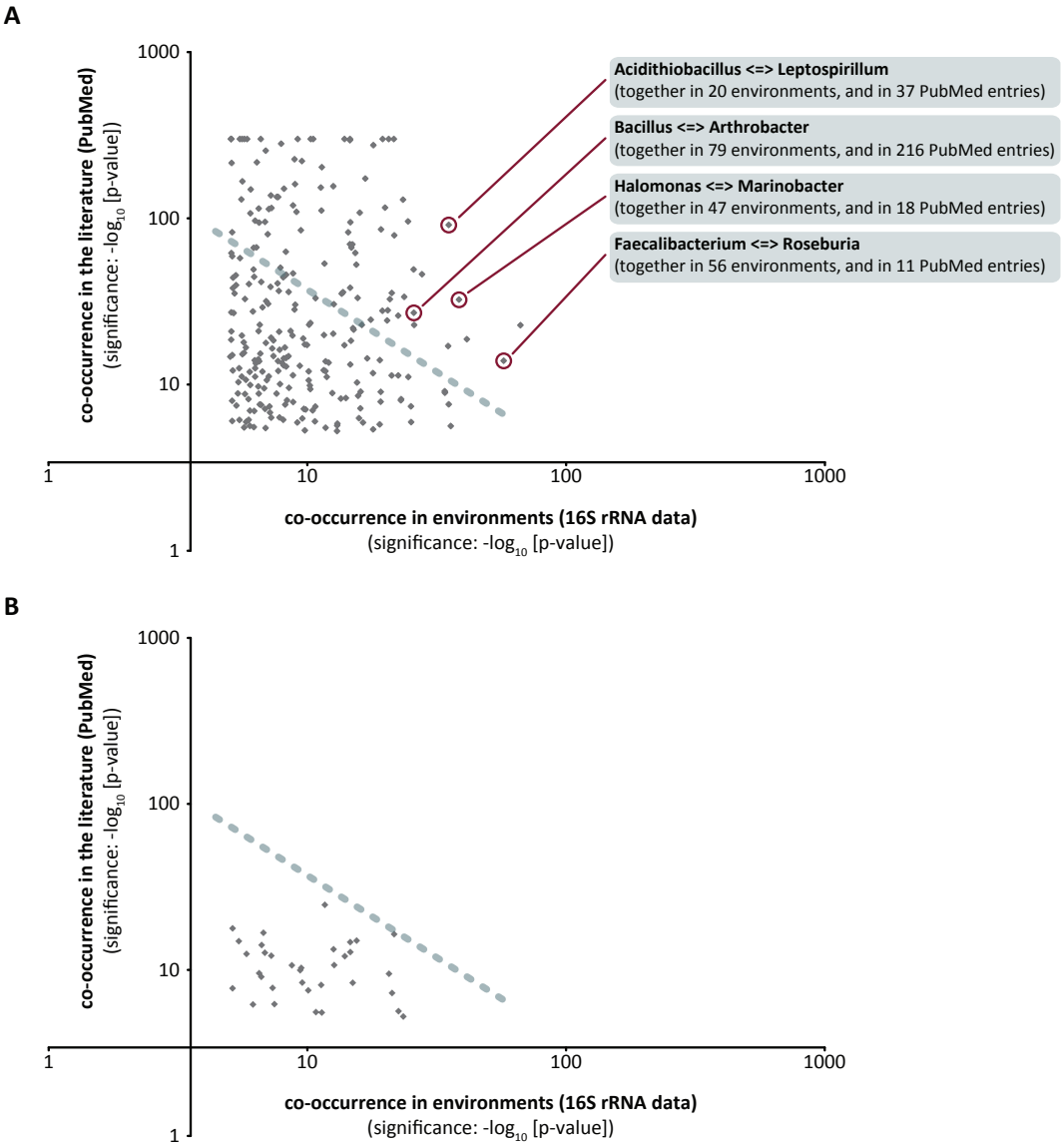


Figure S5: Overlap of our associations with previous knowledge as revealed by textmining.
The PubMed server was queried programmatically for all prokaryotic genus names, as annotated in the NCBI taxonomy, using the 'textword' search (tag 'TW'). We excluded the genus 'Escherichia', since E. coli is very widely mentioned, but often only occurs in technical contexts and less often in ecological contexts. We then computed co-occurrence of pairs of genus names in PubMed entries, exactly like we have done for OTUs in environmental samples. The figure illustrates the overlap between both tests. Each dot denotes a pair of genus names, seen as significantly co-occurring both in PubMed and in our 16S rRNA data. Panel B shows the exact same plot, but after a conservative randomization of PubMed: each entry was randomly assigned to exactly the same number of lineages as it had before the randomization, and each lineage to the same number of PubMed entries. The dotted line, in both plots, represents an arbitrary cutoff, below which all associations are assumed to be random artifacts arising from the skewed size distributions.

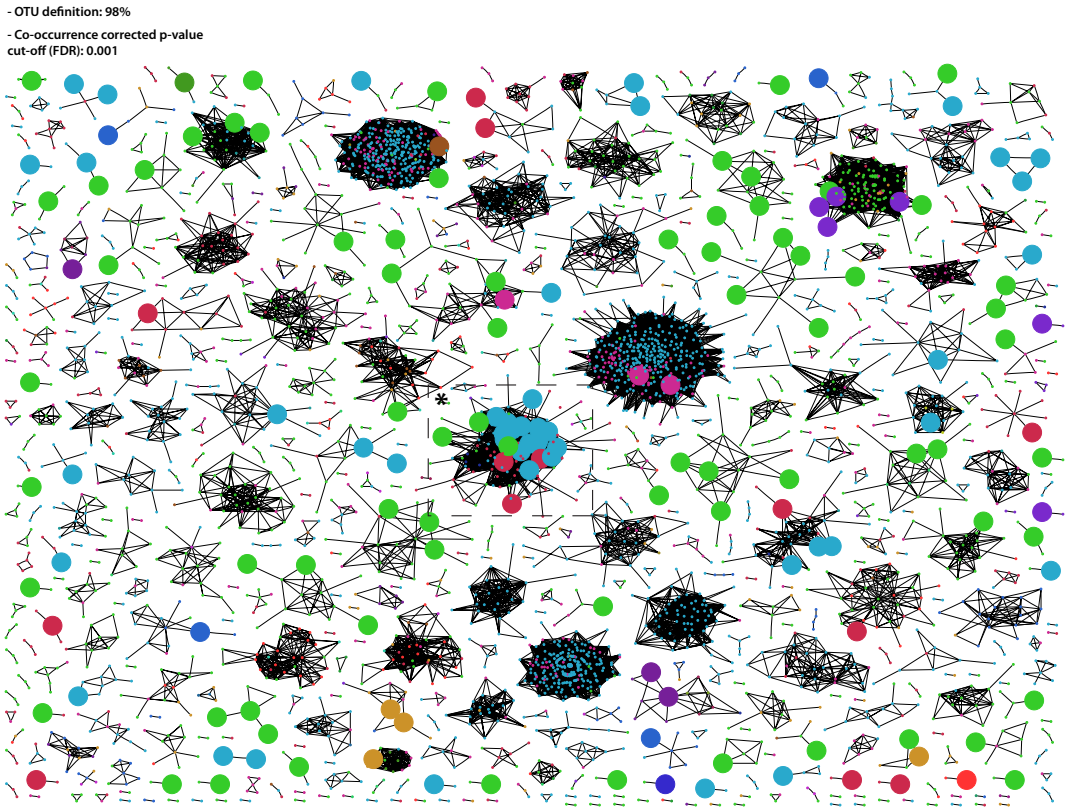


Figure S6: Network of coexisting microbial lineages, genome mappings emphasized. Similar to Figure 2 in the main text, this figure shows microbial lineages that have been found to co-occur (albeit using an OTU definition at 98% sequence identity cutoff); the network has again been partitioned into modules. Here, large nodes denote lineages to which a completely sequenced genome has been mapped. * Note the highly connected module in the center; it contains numerous mapped genomes. In order to test whether this module dominates the signals described in Figure 4 of the main text, all lineages in this module have been omitted, as a control, in the analysis shown in Figure S8.

OTU definition: 97%

Stratified samples	samples <6 seqs	samples >6, <20 seqs	samples >20 seqs
number of co-occurrence tests	561,270	2,011,015	12,199,330
number of coexisting OTUs pairs (FDR=0.001)	1	55	51,161
number of co-occurring OTUs with mapped genome	0	8	81
Coexisting genomes pairs (FDR=0.001)	0	4	169

Down sampling	max 50 seqs/sample	max 100 seqs/sample	max 500 seqs/sample
number of co-occurrence tests	10,353,525	11,623,431	12,502,500
number of coexisting OTUs pairs (FDR=0.001)	1,478	5,204	36,653
number of co-occurring OTUs with mapped genome	59	74	104
Coexisting genomes pairs (FDR=0.001)	54	74	183

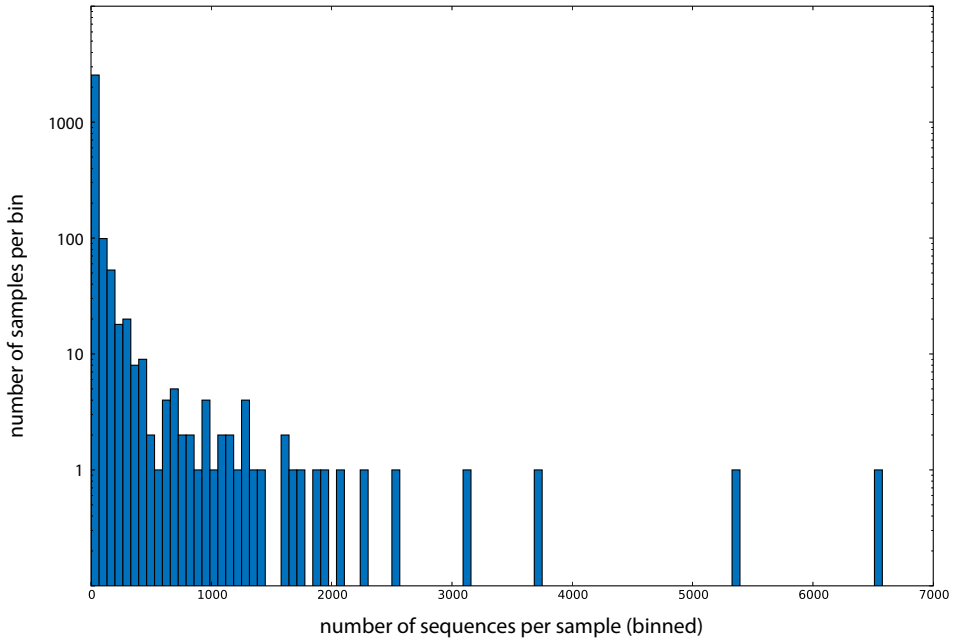


Figure S7: Stratifying environmental samples by sampling depth, results of downsampling, and sample size distributions Environmental sampling of 16S rRNA genes is done at widely varying sampling depths (see bottom part of this figure for the size distribution of samples).
To assess the consequences of this, we subdivided samples into three distinct size bins (top half of table), and compute co-occurrence statistics independently for each bin. As expected, this strongly lowered the number of associations that can be retrieved. Larger samples clearly contribute more signal than smaller samples, but, importantly, none of the bins alone is responsible for the full signal. The full signal (see Table 1 in the manuscript) requires samples from all three size bins. When down-sampling large datasets (bottom half of table), the number of co-existing lineages to which genomes can be mapped does not fall very rapidly: about a third of genome-genome pairs can still be retrieved even when capping all samples at maximally 100 sequence entries.

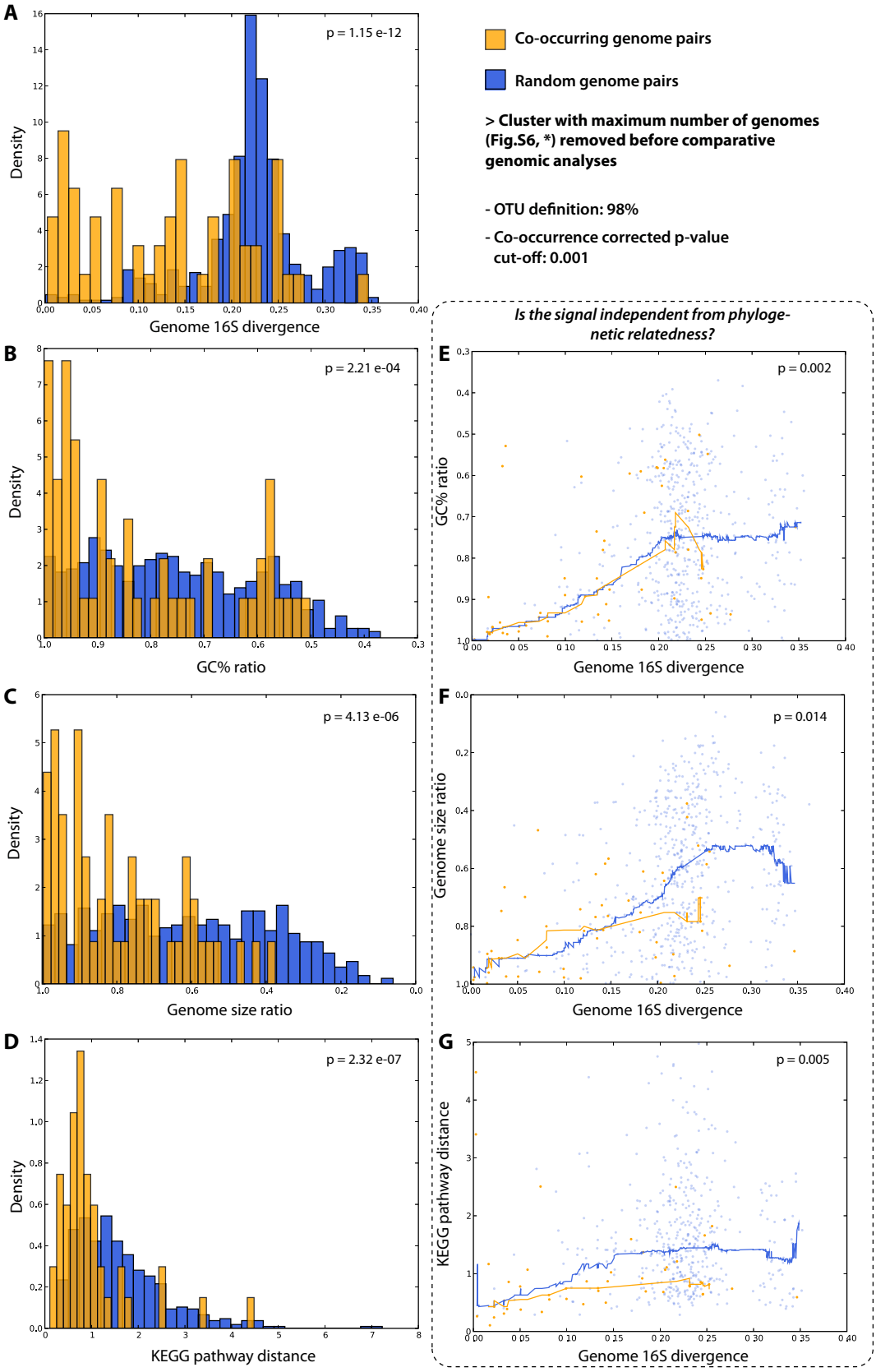


Figure S8: Similarities of genomic features among coexisting lineages do not arise merely from one well-populated cluster.
Here, we repeat all plots shown in Figure 4 of the main text, with the sole difference that we are leaving out lineages mapping to the single module with the largest number of successfully mapped genomes (marked with an asterisk in Figure S6). Note that all reported shifts remain statistically significant.

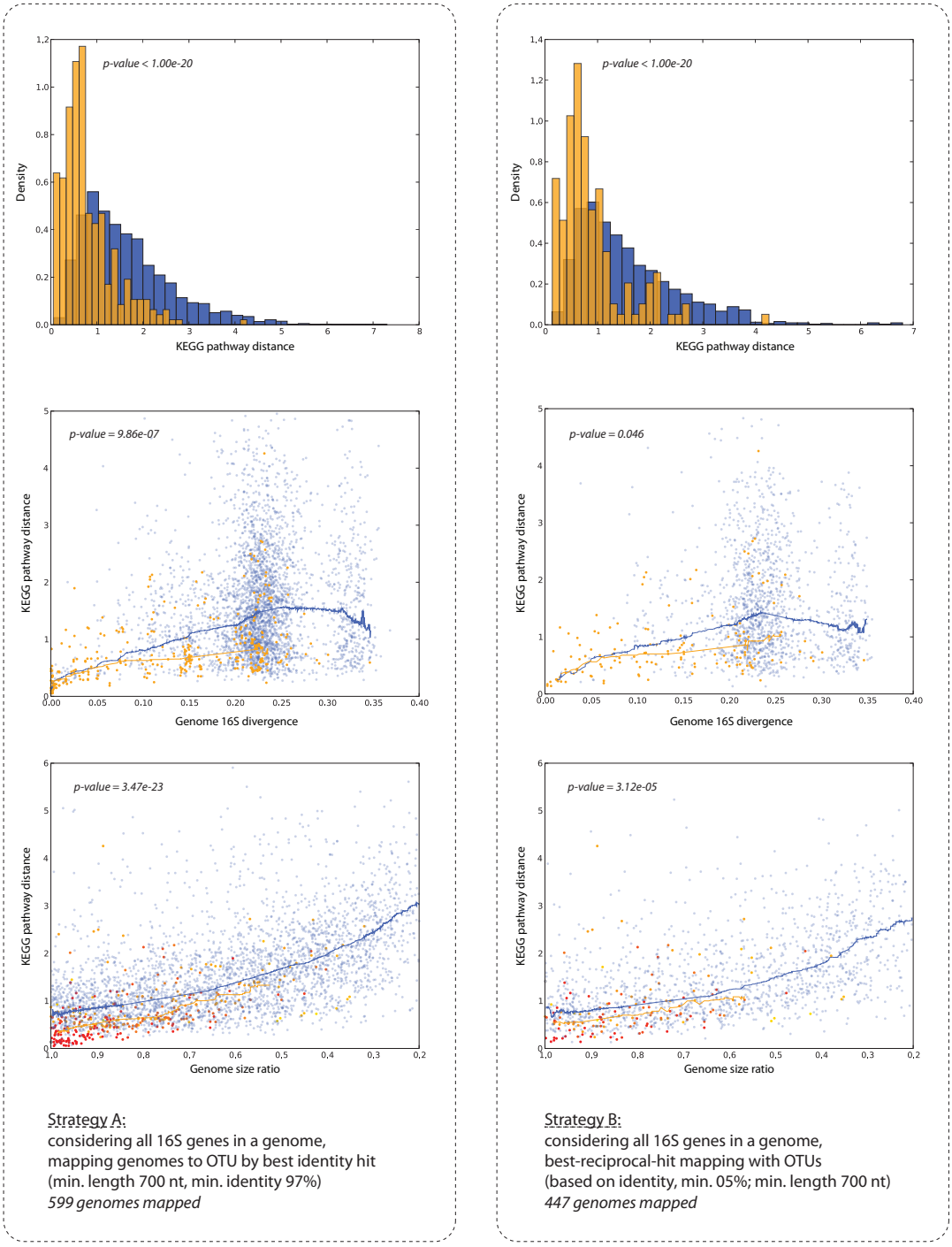


Figure S9: Testing two alternative strategies for genome mapping.
For Figures 4 and 5 of the manuscript, we mapped available, fully sequenced genomes to environmental OTUs – by choosing the longest 16S gene annotated in each genome, and mapping it to available OTUs by 'best hit' in BLAST searches. We also tried two alternative mapping strategies, and reproduce here the results of Figures 4D, 4G, and 5. As can be seen, the quantitative outcome and conclusions remain the same, irrespective of the mapping strategy used.

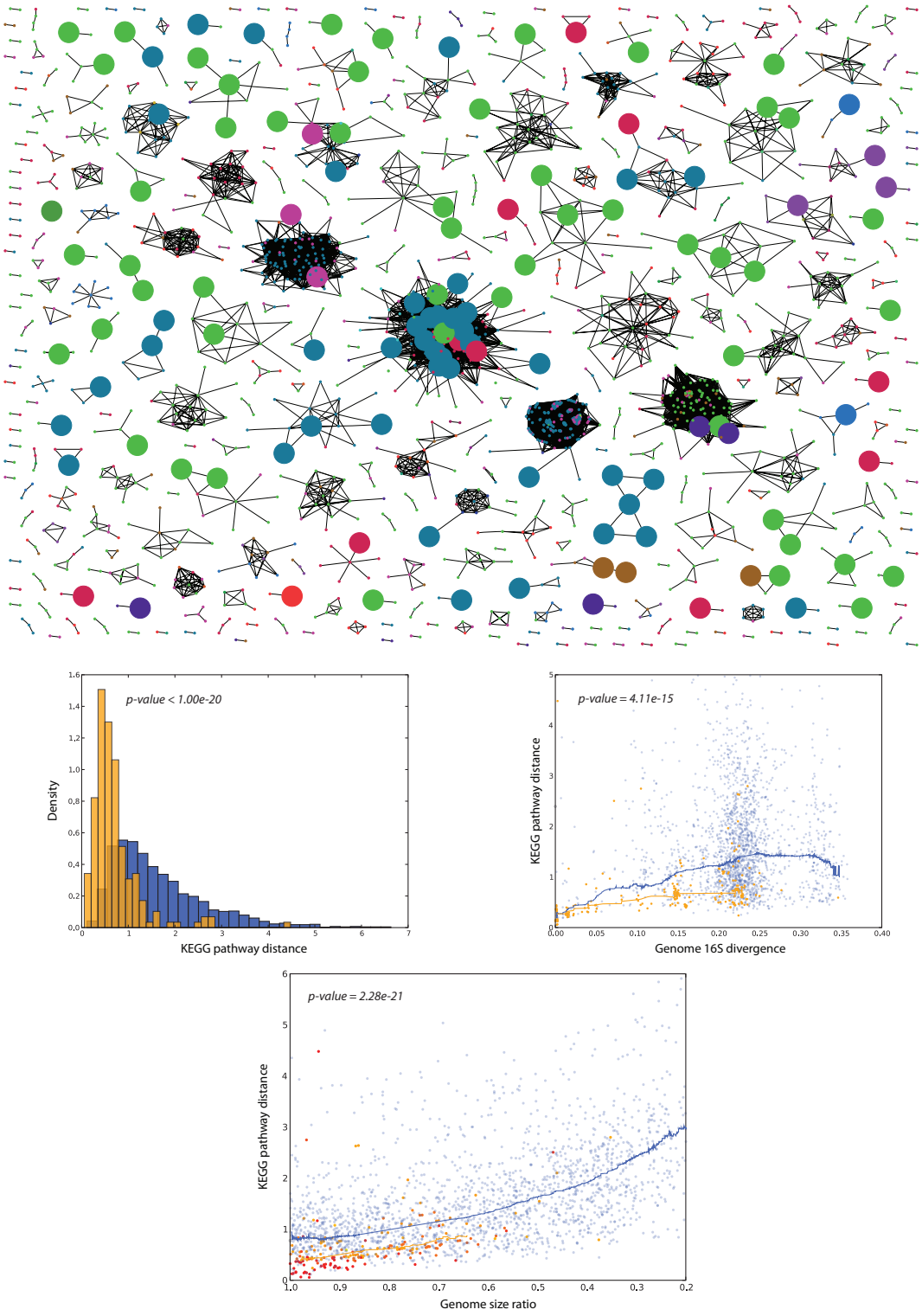


Figure S10: Removing samples related to the mammalian gut.
Microbes in the mammalian gut have been sampled quite extensively; this could potentially lead to distorted results. Therefore, we have tested the effect of removing all gut-related samples: we computed associations separately for all samples that do not have any of the following words in the 'isolation source' description: gut, feces, faeces, fecal, cecal, stool, intestine, intestinal, rumen, colon. The top panel of the figure shows the resulting clustered association network – when comparing this network to the full network in Figure S6, it becomes evident that the gut samples normally result in a number of large, well-connected clusters. Remarkably, these have very few complete genomes mapped to them, which is why our genome-related observations (Figures 4 and 5 in the manuscript), remain essentially unchanged (bottom panels, replicating Figures 4D, 4G, and 5).

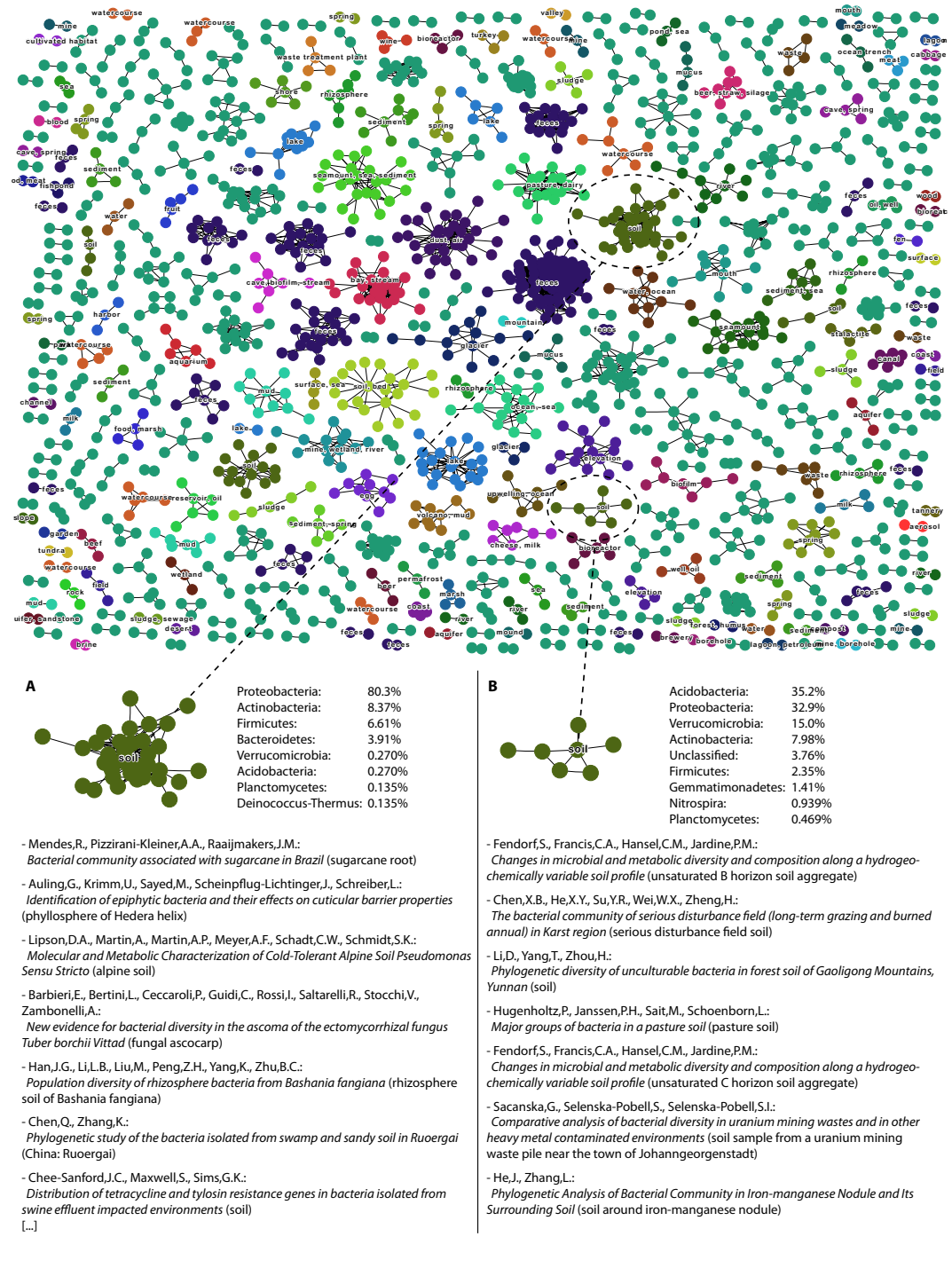


Figure S11: Reverse co-occurrence analysis: objective grouping of sampling events. The figure describes the results of a reverse co-occurrence analysis; instead of grouping microbial lineages by the sampling sites in which they co-occur, here we group sampling events by the microbial lineages that they share. We again perform automatic keyword mapping, and assign distinct colors for each keyword (or group of keywords). Notice how this leads to an objective partitioning of sampling events into habitats. As an example, two clusters are highlighted, which are both annotated with the keyword "soil", but have remarkably different average microbial compositions. The sampling events defining these two modules are indicated; note that the list is truncated for the larger of the two modules.

Part III

DISCUSSION AND PERSPECTIVES

7 DISCUSSION

7.1 HOSTS AND ASSOCIATED MICROBIOTA

Host-microbial symbioses (mutualistic, commensal, or parasitic) occur throughout the phylogeny of animals[98] and also on and within plants. The complexity of communities involved in beneficial associations ranges from low diversity, mostly in invertebrates, to high diversity, for example in vertebrates.

The human intestinal microbiota encompasses roughly ten times more cells than the total number of our somatic and germ cells[4], thus constituting a 'microbial organ' which provides us with additional genetic and metabolic capabilities (e. g., the ability to harvest otherwise inaccessible nutrients). It also plays a crucial role in human health by protecting us from colonization by pathogens. The present situation is thought to be the result of a complex and long co-evolution between host and microbes that led to a mutually beneficial association. It is even hypothesized that the memory-based immune system in vertebrates may have evolved from the necessity to recognize and manage complex communities of microbes[66]. Remarkably, it is also recognized that symbionts can play a crucial role in regulating the development of their host[110].

Although the human intestine is colonized by an enormous population of bacteria, it is dominated by relatively few divisions that are highly diverse at the strain/subspecies level[4]. This can be explained by the fact that intestinal symbionts are selected to be effective consumers of available resources. At the same time, this process also benefits the host because resource competition provides an additional barrier to colonization by potential pathogens. Our recent published work described in Chapter 4 confirms this observation, but at the same time uncovers a potentially more complex mechanism of colonization resistance in which the interactions among microbes play an important role as well. We could show that the actual composition of the community is to some extent predictive of the colonization success by additional cells, be they commensals or pathogens – we observed that colonization is more successful when species that are closely related to the invading species are already present in the gut. This finding motivates research work focusing on the complex interactions and communication processes among bacteria that allow the maintenance (or not) of the equilibrium of our commensal gut microbiota.

7.2 COEXISTENCE, HORIZONTAL GENE TRANSFER AND ADAPTATION

Bacteria and archaea reproduce asexually, but are nevertheless able to exchange DNA elements, and this mechanism is thought to be a major force of evolution and adaptation within these domains of life[40]. The process of HGT can occur between individuals

of the same species, between closely related species and also between distant species – for example between archaea and bacteria[71]. It can be mediated through three different mechanisms: transformation (naked DNA uptake), transduction (via bacteriophages) and conjugation (exchange of plasmids through specialized structures), and constitutes a significant mechanism of innovation[74]. A variety of mobile DNA elements, associated with one or several specific modes of transmission, can often be found within or adjacent to the transferred material. Homologous recombination can mediate the exchange of genetic information between closely related organisms and this allows the fixation of novel, advantageous functions among the population[48,36]. Via illegitimate recombination, HGT can also lead to the dispersal of novel DNA into completely unrelated lineages, and confer new specific abilities to the acceptor. Moreover, such newly acquired capabilities can then also be distributed among closely related organisms via homologous recombination[60]. Transferred elements may be potential markers for the metabolic state of a given habitat; some of them probably code for essential functions required by population members to adapt their metabolism in response to environmental fluctuations or external stimuli. Organisms that have newly entered a habitat can thus exploit novel metabolic functions, thereby facilitating colonization or adaptation to the novel ecological niche[59]. Identification and survey of these processes by the interpretation of metagenomics data should increase our knowledge of mechanisms leading to prokaryotic diversity, adaptation and evolution.
A recent study by Nogueira et al. [73] illustrates well the importance of the mechanism of HGT in the context of microbial interactions in natural communities, by examining the cooperative use of proteins as a model system. They defined secreted proteins (the *secretome*) as 'public goods' since multiple members of a community can benefit from them. They propose that the mechanism of HGT allows cooperating organisms to convert 'cheaters' (using but not producing the 'public goods') into well-behaved community members, by increasing the relatedness of group members. The evidences stems from the analysis of 20 complete genome sequences from strains of *Escherichia coli*. The genes encoding secreted products seem to be lost and regained much more often than genes encoding cytoplasmic proteins (there is a significant association between genes encoding secreted proteins and integrases), enabling the community members to share the genes encoding the cooperative behavior. The authors also detected an association between genes encoding secreted proteins and those belonging to addiction modules, therefore enforcing cooperation by preventing a sudden loss of social genes along with the addiction module, which would result in cell death. Interestingly, these results illustrate the possible role of HGT in the evolution of cooperation among bacteria.

Horizontal transfer of life-style genes in the environment could also be correlated with notable co-occurrences of specific organismal lineages. Lineages that are consistently found together in environments may have a higher likelihood of exchanging genes or operons, and this may give another indication on how microorganisms cooperate, partly taking advantages of these mechanisms of transfer. There might be a kind of common altruistic interest (beneficial symbiosis for both partners) for both populations (genes and species) to support each other (is it possible to demonstrate any stable balance between community members and the genetic material

they exchanged or share?). Preferential coexistence between microbes would increase the amount of exchanged or shared DNA, and these partnerships would therefore increase the evolution and adaptation of the community as a whole[13]. Environmental preferences at the levels of communities, species and genes might also be characterized using this approach and provide more details about how genes and members of a community are in constant competition with each other, or in balance, among the different sampled environments. Finally, assessing the diversity of species and genes in consortia via the analysis of environmental genome sequences could provide insights into ecology. From the perspective of the genome and probably more significantly from the metagenome, DNA sequences reflect the real dynamicity and evolutionary capabilities of organisms in their environment[75]. Species diversity within communities and genetic diversity within populations are hypothesized to co-vary, because of local characteristics that influence the two levels of diversity via parallel processes, or because of direct effects of one level of diversity on the other[119]. The correlation between discoveries at these two levels (organisms and genes) at which natural selection can operate might help us to better understand how a population structure is maintained and stabilized during the course of evolution.

8 PERSPECTIVES

8.1 SPECIFIC TRACES OF NICHE ADAPTATION

Which classes of genes are preferably transferred into and fixed in microbial populations? Are these genes exhibiting a reduced diversity because of selective sweeps? Are biological functions of these genes related with the niche/environment in which they have been identified? Do we already know the majority of such genes, or do they represent the rare, unknown 'tail-end' of the gene frequency distribution?

To tackle these questions, research focus could be oriented on low abundance genes, particularly those that are showing signs of frequent exchange or sharing between bacterial species. These coding modules (or functional entities) might be specific and unique by their way of inheritance. It might be possible to detect traces of homologous recombination acting as a mechanism to lower the diversity of these genes, in contrast to the higher diversity of 'normal' genes which are almost always transmitted vertically. This observation would confirm the influence of strong natural selection created by the environment on coding sequences that play an important role for the maintenance or resistance of organisms in their habitats. The search for such a type of 'reduced diversity' genes could be performed not only among environmental genome sequences, but also in complete genome sequences, where it should be possible to detect them. These genes should be overall rare, have a patchy and non-standard distribution across phyla, but at the same time exhibit lower-than-usual molecular diversity.

One possible approach to measure diversity and to identify entire groups of such genes is to use operon neighborhood information - identifying entire life-style 'modules' specific to an environment, and potentially transferred jointly during HGT events. Operon structure and conservation can be a powerful tool to better understand the dynamics of such genetic elements within a natural community.

We hypothesize that at the low-abundance, 'long tail' end of gene families, many molecular functions can be found that are characterized by frequent horizontal transfers and decreased molecular diversity. Identifying these genes may lead to a better understanding on how bacterial species are able to adapt to diverse habitats, as HGT has been proposed to be responsible for the adaptation of bacteria and archaea to extreme environments[71,63]. Previously, mobile DNA elements such as insertion sequence (IS) have been shown to undergo episodes of positive selection, for example in the genome of a cyanobacterium[68]. Can we identify low-diversity mobile elements or specific 'life-style' genes in metagenomics data? And are these elements supporting an important ecological function to stabilize certain species or a community in its environment? Is there any correlation between genes with reduced diversity and the frequency of HGT? Can selfish interests of mobile DNA elements, exchanged between community members, lead to beneficial outcomes for a community?

8.2 PROTEIN-DNA-GENOME FRAGMENT RECRUITMENT

The fast development of high-throughput sequencing enables the complete sequencing of cultivated microorganisms of interest[133]. This is relevant because reference genomes are indispensable as a basis for metagenomics and functional genomics analyses, setting up a basic and trustable knowledge to extend discoveries. As demonstrated in our analysis in Chapter 5, they allow the automatic functional annotation of omics data but also the assignment to known phylogenetic taxa. The National Institutes of Health has launched an initiative that focuses on describing the diversity of microbial species that are associated with health and disease, and the first phase of this initiative actually includes the sequencing of hundreds of microbial reference genomes[72].

The method we developed for the integration of both genomic and proteomic data, presented in Chapter 5, can be improved. In our approach, the direct recruitment of peptides onto a reference genome might be too stringent for analyzing complex microbial communities because a single amino-acid mismatch between a peptide and a protein from a reference genome can preclude the recruitment. An alternative strategy, depicted in Figure 12, can circumvent this problem. The first step would be peptide searches using a database of open reading frames (ORFs) predicted directly from the metagenome raw reads (without assembly); this approach is more powerful for protein identification because the metagenome represents the true coding potential of the community. The second step is the recruitment of the metagenome reads (whether or not they carry peptides) onto a selected set of complete reference genomes using the program basic local alignment search tool (BLAST). Obviously, the ideal setup for this approach requires the availability of both genomic and proteomic information extracted from the very same samples. Considering taxonomy assignment, this method allows the direct inference of community members' taxonomy, using genomic information, at various levels of resolution by controlling the sequence identity cut-off of recruited reads. This is convenient, because the species definition concept is automatically transformed into an ecosystem species concept.

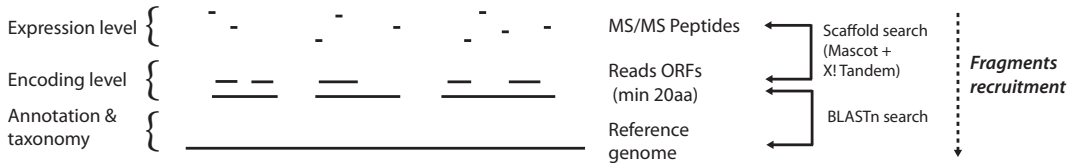


Figure 12: The Protein-DNA-Genome fragment recruitment approach.
Schematic representation of our updated fragment recruitment approach. Peptides are searched against a read-ORF database using protein identification softwares (to match mass spectra with peptide sequences). Metagenome DNA reads are searched against a database of complete reference genomes using BLAST. This method allows to link protein information to the coding potential of the members of a microbial community.

The recruitment of fragments onto a reference genome allows at the same time to compute "encoding levels" (DNA coverage) and "expression levels" (peptide coverage) for each gene encoded in the community genomes. Thus, it is possible to infer

expression scores for each gene by calculating the ratio of peptide coverage over reads coverage. The Figure 13 depicts the results of such calculations for the fragment recruitment approach of three different datasets onto a single complete reference genome (*Methylobacterium radiotolerans*) and its eight associated plasmids. Since the reference genomes are generally annotated by various functional classification systems, these calculations enable a functional and metabolic comparison among various microbial communities (different samples) but also among members of a given microbial community, and as well as studying various taxonomy depths in the phylogenetic tree of the reference genomes.

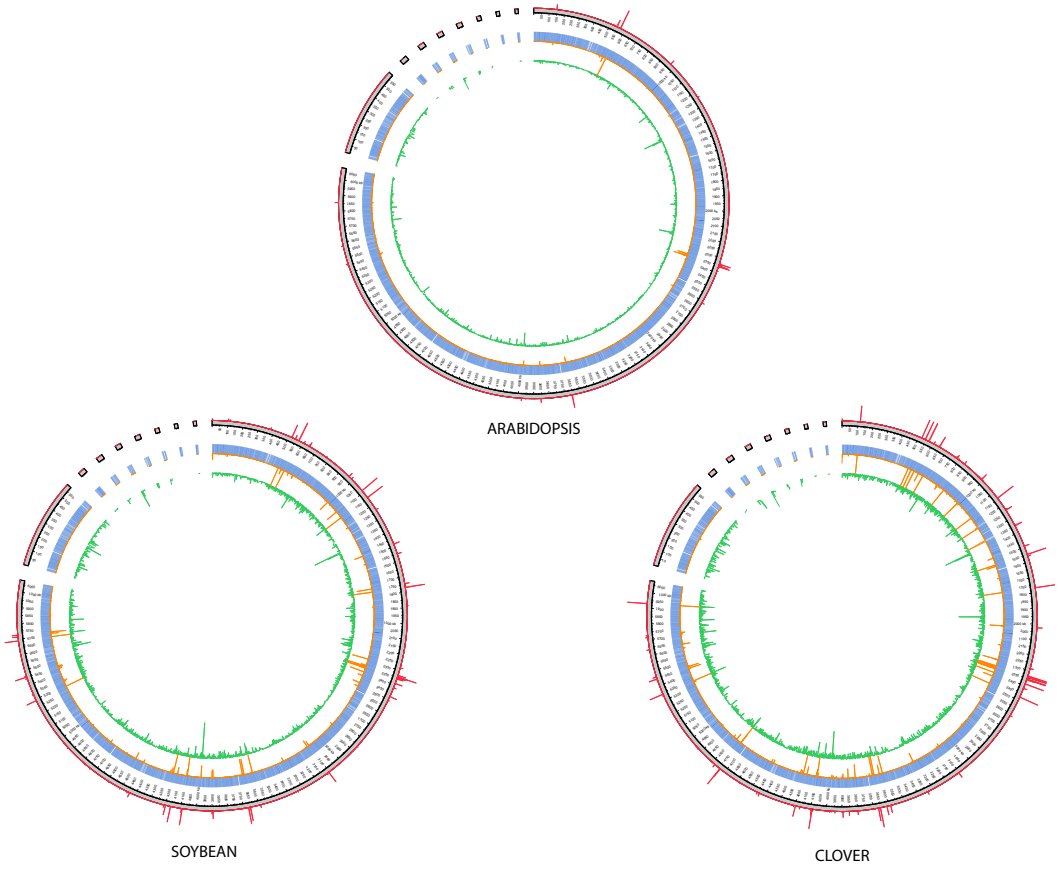


Figure 13: Protein-DNA-Genome fragment recruitment plots for *Methylobacterium radiotolerans*.
The figure (generated using the software circos[57]) illustrates the fragment recruitment of three distinct microbial communities sampled on three different plants (*Arabidopsis thaliana*, soybean and clover) onto the reference genome *Methylobacterium radiotolerans* and its eight plasmids. The encoding levels, the expression levels and the expression scores are depicted in green, orange and red, respectively.

This novel methodology might help us to better understand the physiology of the phyllosphere microbiota and could also uncover mechanisms of bacteria-plant and bacteria-bacteria interactions in the phyllosphere. Importantly, the approach presented here can be applied to a variety of environments colonized by communities ranging from low to high complexity.

8.3 AN ON-LINE RESOURCE FOR MICROBIAL COEXISTENCE

A publicly accessible resource on microbial coexistence could be of great interest to a wide range of researchers in microbiology, ecology, systems biology and also biotechnology. Such a database could be built by mining the publicly available 16S rRNA sequences and their associated meta-information as described in Chapter 6. Importantly, this database could also integrate metagenomics information and associated meta-data. In the future, one can imagine the evolution of this resource into a meta-platform aiming at the integration and mining of all genomic and functional genomic datasets within an ecology-oriented framework.

It is clear that bacterial taxonomic ranks higher than the species level share ecological preferences[84], as we indirectly showed in our microbial coexistence analysis[18]. Therefore, a database as described above could also greatly aid in developing a new classification scheme for microbes by integrating 'omics' data with ecological information.

This resource could also be helpful to microbiologists as a guide to identify ecosystems with potentially interesting microbial consortia. Today, novel techniques and methodologies[76] can help to reveal intimate associations among microbes *in situ*, such as the consortia responsible for the AOM in marine sediments[83,21]. Such associations, given the fact that the strains composing the consortia are known and cultivable, can be extensively studied in the laboratory by "knock-out communities" experiments[56]. Mathematical models and simulations can also contribute to this research by generating hypotheses on the nature of the interactions among the members of a microbial community[47], although interactions among communities counting more than two members seem to be difficult to predict[55].

Part IV

APPENDIX

A

SALMONELLA ENTERICA SEROVAR TYPHIMURIUM EXPLOITS INFLAMMATION TO COMPETE WITH THE INTESTINAL MICROBIOTA

Salmonella enterica Serovar Typhimurium Exploits Inflammation to Compete with the Intestinal Microbiota

Bärbel Stecher[1], Riccardo Robbiani[1☯], Alan W. Walker[2☯], Astrid M. Westendorf[3], Manja Barthel[1], Marcus Kremer[4], Samuel Chaffron[5], Andrew J. Macpherson[6], Jan Buer[3], Julian Parkhill[2], Gordon Dougan[2], Christian von Mering[5], Wolf-Dietrich Hardt[1*]

1 Institute of Microbiology, Swiss Institute of Technology Zurich, Zurich, Switzerland, **2** Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom, **3** Department of Mucosal Immunity, Helmholtz Centre for Infection Research, Braunschweig, Germany, **4** Technical University Munich, Munich, Germany, **5** Bioinformatics Group, Institute of Molecular Biology, University of Zurich, Zurich, Switzerland, **6** Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada

Most mucosal surfaces of the mammalian body are colonized by microbial communities ("microbiota"). A high density of commensal microbiota inhabits the intestine and shields from infection ("colonization resistance"). The virulence strategies allowing enteropathogenic bacteria to successfully compete with the microbiota and overcome colonization resistance are poorly understood. Here, we investigated manipulation of the intestinal microbiota by the enteropathogenic bacterium *Salmonella enterica* subspecies 1 serovar Typhimurium (*S.* Tm) in a mouse colitis model: we found that inflammatory host responses induced by *S.* Tm changed microbiota composition and suppressed its growth. In contrast to wild-type *S.* Tm, an avirulent *invGsseD* mutant failing to trigger colitis was outcompeted by the microbiota. This competitive defect was reverted if inflammation was provided concomitantly by mixed infection with wild-type *S.* Tm or in mice (IL10$^{-/-}$, VILLIN-HA$^{CL4\text{-}CD8}$) with inflammatory bowel disease. Thus, inflammation is necessary and sufficient for overcoming colonization resistance. This reveals a new concept in infectious disease: in contrast to current thinking, inflammation is not always detrimental for the pathogen. Triggering the host's immune defence can shift the balance between the protective microbiota and the pathogen in favour of the pathogen.

Citation: Stecher B, Robbiani R, Walker AW, Westendorf AM, Barthel M, et al. (2007) *Salmonella enterica* serovar Typhimurium exploits inflammation to compete with the intestinal microbiota. PLoS Biol 5(10): e244. doi:10.1371/journal.pbio.0050244

Introduction

The evolution of pathogenic microorganisms has been shaped to a great extent by their interaction with cognate host species. Colonization is the first step of any infection. For enteropathogenic bacteria, this poses a formidable task as the target host organ is already colonized by a dense microbial community, the microflora, or "microbiota". Intestinal colonization by microbiota begins immediately after birth and lasts for life. In a healthy intestine, the microbiota is quite stable, and its gross composition at higher taxonomic levels is similar between individuals, and even between humans and mice [1]. The intestinal ecosystem is shaped by symbiotic interactions between the host and the microbiota. Microbiota composition is influenced by nutrient availability, local pH, and possibly also by the host's immune system [2]. Conversely, the microbiota optimizes nutrient utilization [3,4], and boosts maturation of intestinal tissues and the intestinal immune system [5–7]. In addition, the microbiota provides an efficient barrier against infections ("colonization resistance"), which must be overcome by enteropathogenic bacteria. It is poorly understood how enteropathogens can achieve that task. Here, we used *Salmonella enterica* subspecies 1 serovar Typhimurium (*S.* Tm) and a mouse colitis model to study strategies by which enteropathogenic bacteria break colonization resistance. *S.* Tm infects a broad range of animal species and is a frequent cause of intestinal infections in the human population. The normal murine microbiota provides colonization resistance and prevents intestinal colonization upon oral *S.* Tm infection. Oral treatment with the antibiotic streptomycin (20 mg of streptomycin intragastric [i.g.]) transiently reduces the microbiota by >80% and disrupts colonization resistance for a period of 24 h [8,9]. The residual microbiota re-grows within 2–3 d, and colonization resistance is re-established ([9]; unpublished data). These studies have provided the basis for a "streptomycin mouse model" for *Salmonella* enterocolitis [10]: 1 d after streptomycin treatment, oral infection with *S.* Tm leads to efficient colonization of the murine intestine, especially the cecum and the colon (approximately 10^9 colony-forming units [CFU]/gram; Figures 1A and S1) [8,9,11]. Wild-type *S.* Tm (*S.* Tm^{wt}) triggers pronounced intestinal inflammation (colitis) and colonizes the intestinal

Academic Editor: Matt Waldor, Harvard University, United States of America

Received September 27, 2006; **Accepted** July 16, 2007; **Published** August 28, 2007

Abbreviations: CFU, colony-forming units; FISH, fluorescence in situ hybridization; HE, hematoxylin and eosin; IBD, inflammatory bowel disease; i.g., intragastric; *L. reuteri* RR^{Rif}, rifampicin-resistant variant of *Lactobacillus reuteri* strain RR; mLN, mesenteric lymph nodes; p.i., post-infection; *S.* Tm, *Salmonella enterica* subspecies 1 serovar Typhimurium; *S.* Tm^{avir}, *Salmonella enterica* subspecies 1 serovar Typhimurium Δ*invG sseD::aphT*; *S.* Tm^{wt}, wild-type *Salmonella enterica* subspecies 1 serovar Typhimurium; SPF, specified pathogen free

* To whom correspondence should be addressed. E-mail: hardt@micro.biol.ethz.ch

☯ These authors contributed equally to this work.

Author Summary

A dense microbial community colonizes the intestinal tract of mammals, contributing to health and nutrition and conferring efficient protection against most pathogenic intruders. Intestinal pathogens can overcome this colonization resistance and cause disease; however, the mechanisms used to do this are still elusive. In this study we analyzed intestinal infection by the model pathogen *Salmonella enterica* subspecies 1 serovar Typhimurium (*S.* Tm). We show that the virulent wild-type pathogen overcomes colonization resistance by inducing the host's inflammatory immune response and exploiting it for its purpose. In contrast, an avirulent *Salmonella* mutant defective in triggering inflammation was unable to overcome colonization resistance by itself. Colonization by this mutant was restored if inflammation was provided concomitantly, in mice with inflammatory bowel disease (genetic and inducible) or by co-infection with wild-type *S.* Tm. These findings reveal a previously unrecognized strategy by which pathogenic bacteria overcome colonization resistance: abusing the host's inflammatory immune response to gain an edge against the normal microbial community of the gut. This represents a first step towards unravelling the molecular mechanisms underlying this three-way interaction of host, microbiota, and pathogens.

lumen at high densities over extended periods of time [8,10–12]. This "streptomycin mouse model" can be used to study bacterial virulence factors required for colonization and triggering of intestinal inflammation. For example, *S.* Tm strains lacking the two virulence-associated type III secretion systems (e.g., *S.* Tm Δ*invG sseD::aphT* [*S.* Tm^{avir}] [13]) cannot trigger colitis. In addition, these mutants were found to colonize the murine intestine only transiently [11,13]. The reason for this colonization defect has remained elusive.

To explore this, we analyzed microbiota compostition in *S.* Tm^{wt}– and *S.* Tm^{avir}–infected mice and the role of inflammation for *Salmonella* colonization and competition against the intrinsic microbiota. We found that inflammation shifts the balance between the protective microbiota and the pathogen *S.* Tm in favour of the pathogen. This principle might apply to various other pathogens and therefore constitute a novel paradigm in infectious biology.

Results

S. Tm^{avir} but Not *S.* Tm^{wt} Is Outcompeted by Commensal Microbiota

First, we confirmed the differential colonization efficiency of *S.* Tm^{wt} and *S.* Tm^{avir} in the streptomycin mouse model. Unlike *S.* Tm^{wt}, intestinal *S.* Tm^{avir} colonization levels decreased significantly by day 4 post-infection (p.i.) in a highly reproducible fashion (Figure 1B). This coincided with re-growth of the microbiota as revealed by immunofluorescence microscopy (Figure 1C–1H). By anaerobic culture, DNA isolation, and 16S rRNA gene sequencing, high densities of characteristic members of the intestinal microbiota (*Clostridium* spp., *Bacteroides* spp., and *Lactobacillus* spp. [14]) were found in *S.* Tm^{avir}–infected, but not in *S.* Tm^{wt}–infected, animals at day 4 p.i. (Table 1). Both the *S.* Tm/microbiota ratio and the composition of the microbiota itself differed between mice infected with *S.* Tm^{avir} and *S.* Tm^{wt}. These data demonstrated that residual microbiota surviving the streptomycin treatment can re-grow, outcompete *S.* Tm^{avir}, and thereby re-establish colonization resistance. In contrast, *S.* Tm^{wt} can suppress re-growth of the residual microbiota. Therefore, the streptomycin mouse model allows study of the principal mechanisms by which enteropathogens manipulate the intestinal ecosystem.

S. Tm^{wt} Alters Composition of the Microbiota in the Streptomycin Mouse Model

To better characterize the effect of *S.* Tm on microbiota composition, we employed 16S rRNA gene sequencing (see Materials and Methods). This method allows a quantitative comparison of microbial communities, including bacterial species that cannot be cultivated in vitro. The analysis comprised five different groups of mice and addressed the effect of the streptomycin pretreatment per se as well as the effect of *S.* Tm^{avir} and *S.* Tm^{wt} infection on microbiota composition (Figure 2).

In line with published data, a large fraction of the murine microbiota in unmanipulated mice belonged to either the Firmicutes (including *Clostridium* spp. and *Lactobacillus* spp.; 39% ± 10%) or the Bacteroidales (53% ± 13%; Figure 2) [1,15–17]. Streptomycin treatment reduced the global density of the microbiota by approximately 90% (Figure 2; see also Figure 1C and 1D) and changed its relative composition (Figure 2A and 2B; Table 2). The composition of the remaining microbiota varied substantially between individual members of this group (Figure 2B). Most likely, this is attributable to the unstable situation created by the antibiotic and may arise from slight animal-to-animal variations in the timing or speed of the gut passage of the antibiotic and/or from species-specific differences in antibiotic susceptibility and rate of re-growth.

Five days after the antibiotic treatment, the microbiota had re-grown to normal density and microbiota composition, at least at the phylum level (Figure 2A and 2B; Table 2; $p = 0.35078$). Infection with *S.* Tm^{avir} did not interfere detectably with re-growth of the normal microbiota in the streptomycin-pretreated mouse model (Figure 2B; Table 2).

In contrast, *S.* Tm^{wt} significantly altered the cecal microbiota composition (Figure 2A and 2B; Table 2; $p < 0.00001$). Proteobacterial 16S rRNA gene sequences represented >90% of all sequences, and *Salmonella* spp. generally represented the most prominent (up to 100%) proteobacterial species in the *S.* Tm^{wt}–infected animals. These observations were confirmed by fluorescence in situ hybridization (FISH) of fixed cecal content (Figure S2). This demonstrates that *S.* Tm^{wt} interferes with microbiota re-growth and represents the predominant species at day 4 p.i.

It should be noted that other proteobacterial species (e.g., *Escherichia coli*) were also present in significant numbers in the cecum of most *S.* Tm^{wt}–infected animals (Figure 2A). These proteobacterial strains are low abundance members of the normal gut microbiota of our mouse colony ($<10^7$ CFU/g of cecal content). In many mice the proportion of these commensal proteobacterial species increased concomitant with the *S.* Tm^{wt} infection. This suggests that other bacterial species closely related to *S.* Tm may also be able to benefit from the *S.* Tm^{wt}–triggered inflammation. Further work will be required to address this issue.

The observed changes in microbiota growth in *S.* Tm^{wt}–infected mice were verified in a competitive infection experiment with a specific member of the microbiota. For this purpose we selected a rifampicin-resistant variant of

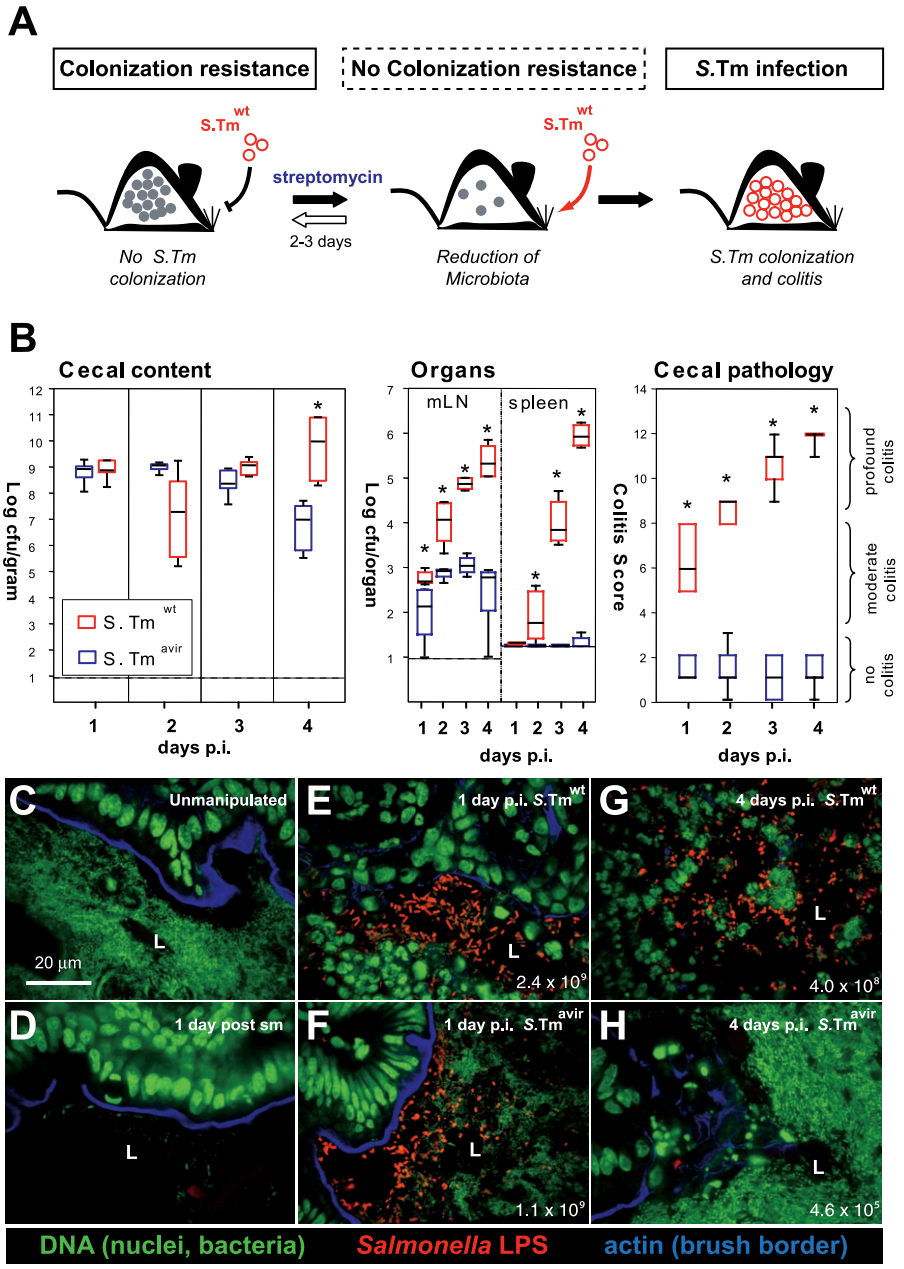


Figure 1. Microbiota Outcompete *S.* Tm^{avir} but not *S.* Tm^{wt}

(A) Streptomycin-treated mouse model. The antibiotic transiently reduces the microbiota (grey circles) in the lumen of the large intestine, reduces colonization resistance, and allows colonization and induction of colitis by *S.* Tm^{wt}.

(B) Streptomycin-treated C57BL/6 mice ($n = 7$ per group) were infected with *S.* Tm^{avir} (blue) or *S.* Tm^{wt} (red; 5×10^7 CFU i.g.). At indicated time points mice were sacrificed, *S.* Tm loads were determined in cecal content, mLN, and spleen, and cecal pathology was scored. Detection limits (dotted lines): cecal content, 10 CFU/g; mLN, 10 CFU/organ; spleen, 20 CFU/organ. *, $p \leq 0.05$; statistically significant difference between *S.* Tm^{avir} and *S.* Tm^{wt}. Boxes indicate 25th and 75th percentiles, black bars indicate medians, and whiskers indicate data ranges.

(C–H) Representative confocal fluorescence microscopy images of cecum tissue sections from the mice shown in (B). Nuclei and bacterial DNA are stained by Sytox green (green), the epithelial brush border actin by Alexa-647-phalloidin (blue), and extracellular *S.* Tm in the intestinal lumen by anti–*S.* Tm LPS antiserum (red). Normal microbiota in unmanipulated mice (C), microbiota 1 d after streptomycin (sm) treatment (D), streptomycin-treated mice infected for 1 or 4 d with *S.* Tm^{avir} or *S.* Tm^{wt} (E–H). The *S.* Tm colonization levels are indicated (CFU/g); L, cecum lumen.

doi:10.1371/journal.pbio.0050244.g001

Lactobacillus reuteri strain RR (*L. reuteri* RR^{Rif}). This strain was isolated as a commensal from our mouse colony. Streptomycin-treated mice were infected i.g. with either *S.* Tm^{wt} or *S.* Tm^{avir} (5×10^7 CFU i.g.) and gavaged 1 d p.i. with *L. reuteri* RR^{Rif} (8×10^6 CFU i.g.). *L. reuteri* RR^{Rif} colonized the *S.* Tm^{avir}–infected mice at levels of 10^5–10^6 CFU/g of intestinal content. In *S.* Tm^{wt}–infected mice, similar *L. reuteri* RR^{Rif} colonization levels were observed at day 2 p.i., but colonization levels declined below the detection limit by day 4 p.i. ($p = 0.008$; Figure 3). Thus, alteration of microbiota composition by *S.* Tm^{wt} can be demonstrated at the level of a single bacterial strain.

Table 1. Bacterial Genera Recovered by Anaerobic Culture from *S.* Tm Infected Mice

Taxonomy according to 16S rRNA Gene Sequence		Day 4 after *S.* Tm^{avir} Infection		*Day 4 after S. Tm^{wt} Infection*	
Genus	**Phylum**	**Percent of Colonies Analyzed[a]**	**CFU/gram**	**Percent of Colonies Analyzed[b]**	**CFU/gram**
Salmonella spp.	Proteobacteria	<1%	6.56×10^{06}	87.5%	6.40×10^{09}
Enterococcus spp.	Firmicutes	<1%	n.d.	7.9%	5.78×10^{08}
Lactobacillus spp.	Firmicutes	29.3%	2.24×10^{09}	4.6%	3.33×10^{08}
Clostridium spp.	Firmicutes	28.3%	1.71×10^{09}	<1%	$<3 \times 10^{07}$
Erysipelothrix spp.	Firmicutes	<1%	3.60×10^{07}	<1%	$<3 \times 10^{07}$
Bacteroides spp.	Bacteroidetes	41.8%	3.93×10^{09}	<1%	$<3 \times 10^{07}$
Total CFU/gram			7.93×10^{09}		7.31×10^{09}

[a]Total of 1,437 colonies.
[b]Total of 329 colonies (see Materials and Methods).
n.d., not determined.
doi:10.1371/journal.pbio.0050244.t001

Intestinal Inflammation Is Sufficient to Enhance Colonization by *S.* Tm^{avir}

The above findings prompted us to investigate whether there is a cause-and-effect relationship between triggering of inflammation and enhanced colonization by *S.* Tm. In this case one would predict that *S.* Tm^{avir} (which cannot trigger inflammation) competes successfully with the microbiota if inflammation is triggered by other means. Three different experimental approaches lent evidence for this hypothesis:

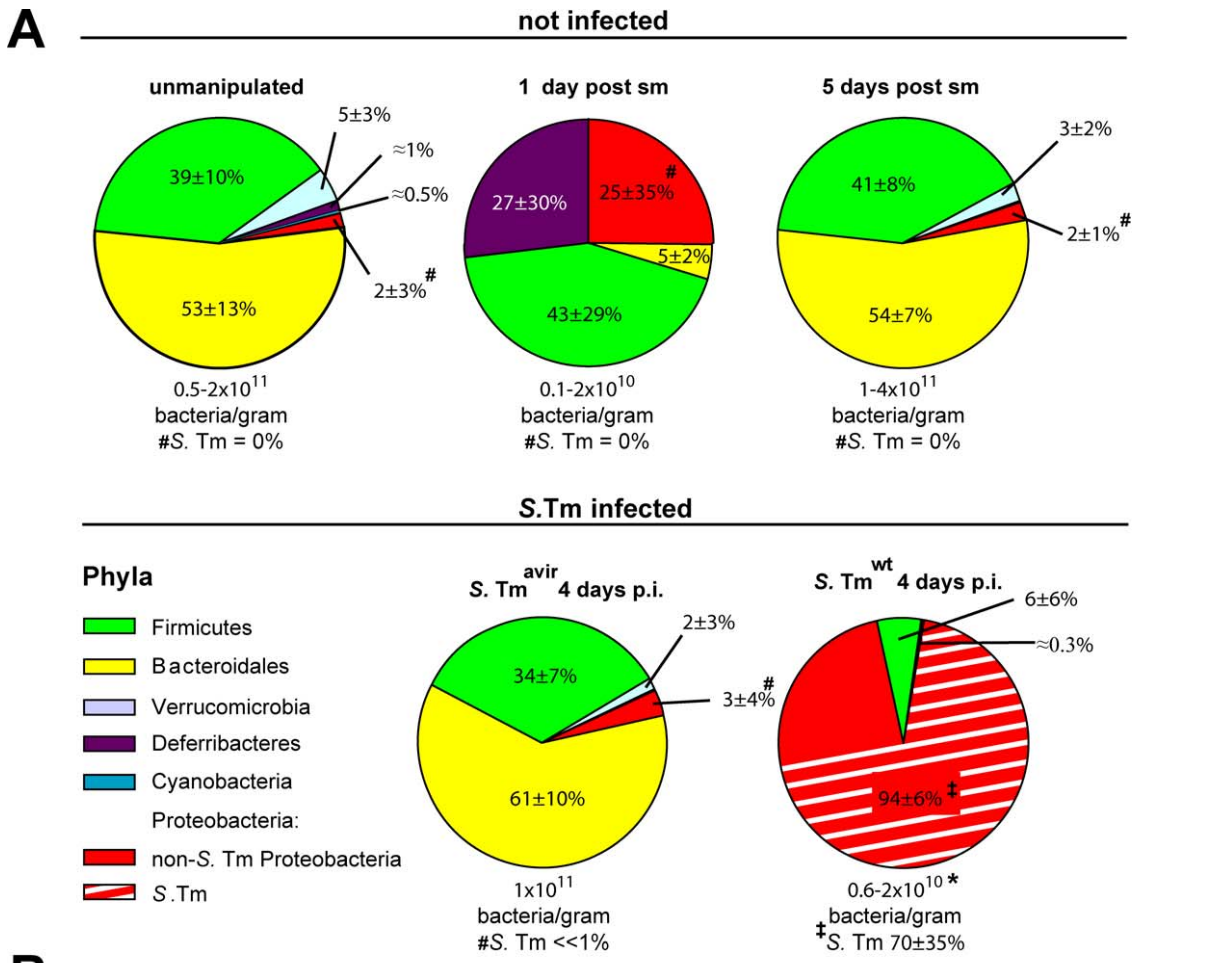
First, we analyzed whether inflammation induced by *S.* Tm^{wt} improved *S.* Tm^{avir} colonization efficiency. Earlier experiments had shown that infections with 1:1 mixtures of *S.* Tm^{wt} and attenuated mutants led to full-blown colitis (Figure 4A and data not shown). Thus, streptomycin-treated mice were infected with a 1:1 mixture of *S.* Tm^{wt} and *S.* Tm^{avir} (a total of 5×10^7 CFU i.g.). Control groups were infected with *S.* Tm^{wt} or *S.* Tm^{avir} only (5×10^7 CFU i.g.; Figure 4A). Pronounced colitis was observed in all animals infected with *S.* Tm^{wt} and the *S.* Tm^{wt}–*S.* Tm^{avir} mixture, but not in animals infected with *S.* Tm^{avir} alone. Furthermore, *S.* Tm^{avir} was severely defective at colonizing lymph nodes and spleen in single and mixed infections. Despite its non-pathogenic phenotype, *S.* Tm^{avir} colonized the cecal lumen up to wild-type levels in mixed infections with *S.* Tm^{wt}. Thus, concomitant colitis created favourable conditions in the intestinal lumen that suppressed microbiota regrowth and rescued *S.* Tm^{avir} colonization in tandem. This was confirmed in long-term infection experiments using 129Sv/Ev mice, which develop a chronic form of colitis (Figures 4B and S3) [12].

Next, we studied whether cecal inflammation per se (in absence of *S.* Tm^{wt}) could enhance *S.* Tm^{avir} colonization. For this purpose we employed knockout mouse models lacking the key anti-inflammatory cytokine IL10. Depending on the exact genetic background and the composition of the microbiota, these animals develop colitis spontaneously earlier (week 6; C3H/HeJBir$^{IL10-/-}$ model [18]) or later in life (week 30–50; C57BL/6$^{IL10-/-}$ model [19]). To test the effect of pre-existing colitis on *S.* Tm^{avir} colonization, groups of 8-wk-old C3H/HeJBir$^{IL10-/-}$ mice and C3H/He control mice were infected (5×10^7 CFU of *S.* Tm^{avir} i.g.; no streptomycin treatment). Fecal shedding (day 1 p.i.), colonization, and colitis (day 2 p.i.) were analyzed. Colonization of the intestinal lumen by *S.* Tm^{avir} was significantly enhanced in mice displaying colitis (day 2 p.i., $p = 0.016$; Figures 5A, S3, and S4). Similar observations were made using the C57BL/6$^{IL10-/-}$ model. In C57BL/6$^{IL10-/-}$ mice, the onset of colitis is quite random and varies anywhere from 30 to 50 wk even between littermates. Accordingly, we infected C57BL/6$^{IL10-/-}$ littermates 30–50 wk of age (5×10^7 CFU of *S.* Tm^{avir} i.g.; no streptomycin treatment). Again, colonization of the intestinal lumen by *S.* Tm^{avir} was enhanced in littermates displaying colitis (day 1 p.i., $p = 0.016$; Figures 5B, S3, and S4). This suggested that inflammation per se can enhance *S.* Tm^{avir} colonization.

To verify this hypothesis we employed the alternative, recently developed VILLIN-HA$^{CL4\text{-}CD8}$ mouse model for T cell–induced colitis [20]. This model employs VILLIN-HA transgenic mice expressing the HA epitope in the gut epithelium and T cells ($CD8^+$; HA-directed α/β T cell receptor; from CL4-TCR transgenic mice) recognizing the

Figure 2. 16S rRNA Gene Sequence Analysis of Microbiota Manipulation by *S.* Tm^{wt} and *S.* Tm^{avir} in the Streptomycin Mouse Model

Cecal contents were recovered from unmanipulated mice, mice at days 1 or 5 after streptomycin treatment (20 mg i.g.), and streptomycin-treated mice 4 d after infection with *S.* Tm^{avir} and *S.* Tm^{wt} (5×10^7 CFU i.g.; all $n = 5$). Total DNA was extracted, and bacterial 16S rRNA genes were PCR-amplified using universal bacterial primers, cloned, and sequenced (approximately 100 sequences per animal; five animals per group; see Materials and Methods).
(A) Pie diagrams showing the microbiota composition at the phylum level. Numbers below the diagrams indicate bacteria/gram cecal content as defined by Sytox green staining. *The lower bacterial density in *S.* Tm^{wt}–infected mice is attributable to a high proportion of cellular debris in the intestinal lumen (see Figure 1G). #In these groups no *Salmonella* 16S rRNA genes were identified. ‡Proteobacterial sequences belonged to *Salmonella (E. coli)* in the following percentages: 91 (1), 15 (70), 87 (11), 55 (38), and 100 (0). See also Table S1.
(B) Visual depiction of the microbiota composition of individual mice. The animals were grouped based on the similarity of their microbiota composition at the phylum level (using the Canberra distance as metric). The resulting groupings are depicted as a dendrogram, and observed phylum counts for each mouse are shown as a heat map (0%–100% of all identified 16S rRNA gene sequences). Labels indicate unique mouse identifier numbers. The experimental groups are indicated. p.sm., post–streptomycin treatment.
doi:10.1371/journal.pbio.0050244.g002


A
not infected
unmanipulated
39±10%
5±3%
≈1%
≈0.5%
53±13%
2±3%#
0.5-2x10^11 bacteria/gram
#S. Tm = 0%
1 day post sm
27±30%
25±35%#
5±2%
43±29%
0.1-2x10^10 bacteria/gram
#S. Tm = 0%
5 days post sm
41±8%
3±2%
2±1%#
54±7%
1-4x10^11 bacteria/gram
#S. Tm = 0%
S.Tm infected
Phyla
Firmicutes
Bacteroidales
Verrucomicrobia
Deferribacteres
Cyanobacteria
Proteobacteria:
non-S. Tm Proteobacteria
S.Tm
S. Tm avir 4 days p.i.
34±7%
2±3%
3±4%#
61±10%
1x10^11 bacteria/gram
#S. Tm <<1%
S. Tm wt 4 days p.i.
6±6%
≈0.3%
94±6%‡
0.6-2x10^10* bacteria/gram
‡S. Tm 70±35%

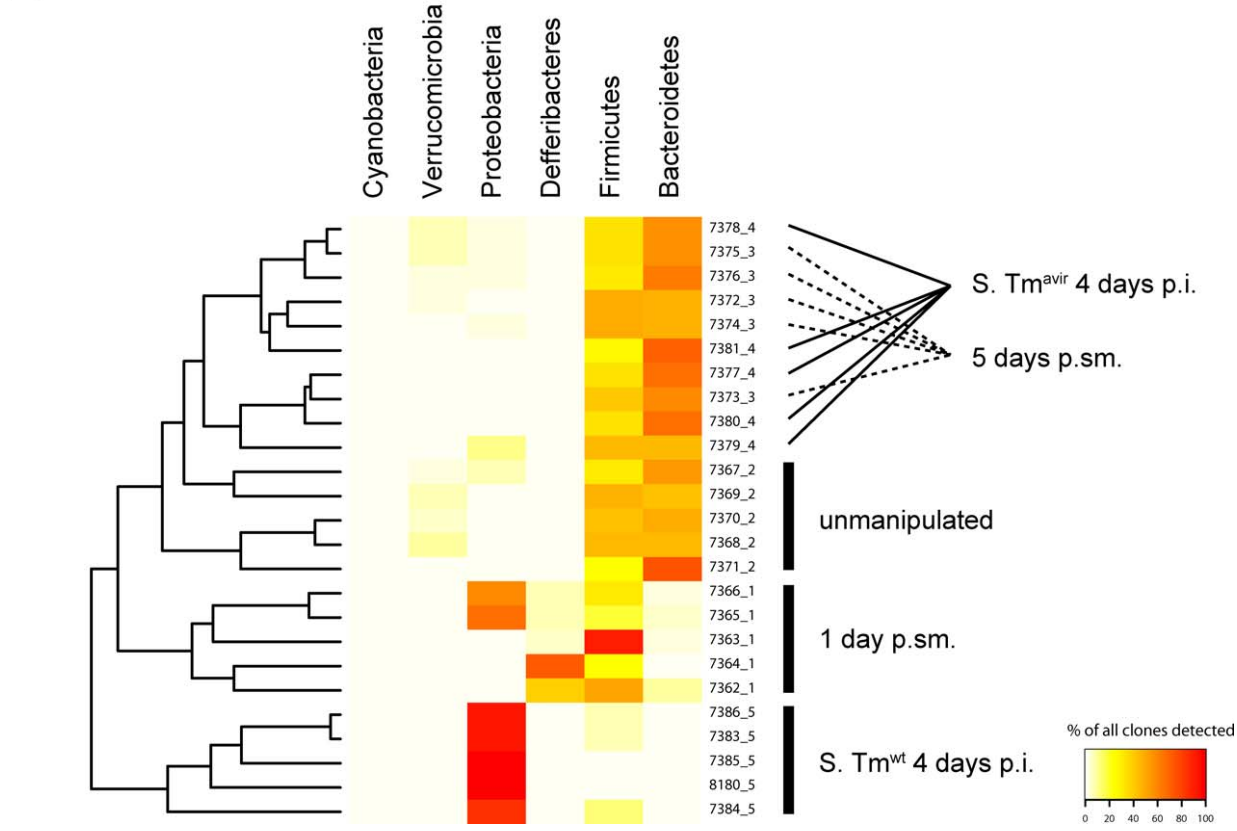

B
Cyanobacteria
Verrucomicrobia
Proteobacteria
Defferibacteres
Firmicutes
Bacteroidetes
7378_4
7375_3
7376_3
7372_3
7374_3
7381_4
7377_4
7373_3
7380_4
7379_4
7367_2
7369_2
7370_2
7368_2
7371_2
7366_1
7365_1
7363_1
7364_1
7362_1
7386_5
7383_5
7385_5
8180_5
7384_5
S. Tm avir 4 days p.i.
5 days p.sm.
unmanipulated
1 day p.sm.
S. Tm wt 4 days p.i.
% of all clones detected
0 20 40 60 80 100

Table 2. Phylum-Level Comparison of Microbiota in Streptomycin-Treated *S.* Tm–Infected Mice from Experiment Described in Figure 2

Group 1	Group 2	*p*-Value	*Difference*
5 d p. sm	Unmanipulated	0.35078	Indistinguishable
4 d p. *S.* Tm^{avir}	5 d p. sm	0.02493	Indistinguishable
4 d p. *S.* Tm^{avir}	Unmanipulated	0.00206	Difference
1 d p. sm	4 d p. *S.* Tm^{avir}	<0.00001	Clear difference
1 d p. sm	4 d p. *S.* Tm^{wt}	<0.00001	Clear difference
1 d p. sm	5 d p. sm	<0.00001	Clear difference
1 d p. sm	Unmanipulated	<0.00001	Clear difference
4 d p. *S.* Tm^{avir}	4 d p. *S.* Tm^{wt}	<0.00001	Clear difference
4 d p. *S.* Tm^{wt}	5 d p. sm	<0.00001	Clear difference
4 d p. *S.* Tm^{wt}	Unmanipulated	<0.00001	Clear difference

$p \geq 0.005$ indicates no significant difference detectable (see Materials and Methods).
p., post; sm, streptomycin.
doi:10.1371/journal.pbio.0050244.t002

HA epitope. Adoptive transfer of these T cells into VILLIN-HA transgenic mice results in severe inflammation of the small and the large intestine at 4–5 d post-transfer (Figure 6A) [20]. This model was of particular interest because intestinal inflammation develops quickly, occurs in the majority of animals, and does not involve i.g. treatment with chemicals that might themselves influence the microbiota–pathogen competition.

To study the impact of inflammation on *S.* Tm^{avir} colonization we infected VILLIN-HA transgenic mice receiving CL4-CD8 T cells and unmanipulated VILLIN-HA control mice. In the unmanipulated VILLIN-HA mice (no T cells transferred), no intestinal inflammation was apparent and *S.* Tm^{avir} colonization efficiency was low (Figure 6B). In contrast, the animals receiving 4×10^6 CL4-CD8 T cells (VILLIN-$HA^{CL4\text{-}CD8}$ mice) developed intestinal inflammation 4 or 5 d after T cell transfer, and *S.* Tm^{avir} efficiently colonized the intestine of these animals ($p = 0.01$; Figure 6B). It should be noted that the initial colonization by *S.* Tm^{avir} was poor (fecal samples at days 2 and 3 after T cell transfer) and that the onset of efficient *S.* Tm^{avir} colonization closely correlated with the onset of the intestinal inflammation (day 4–5 after T cell transfer [20]). At this stage, "*Salmonella*" sequences represented 26%–46% of all bacterial 16S rRNA genes recovered from the cecal contents (Figure 6C). This confirmed that colitis per se creates conditions in the gut skewing the competition between *Salmonella* spp. and the microbiota in favour of the pathogen.

As additional controls, we analyzed the fecal microbiota composition of unmanipulated VILLIN-HA transgenic mice ($n = 4$) and non-infected VILLIN-HA transgenic mice ($n = 2$) at day 4 after CL4-CD8 T cell transfer (Figures 6C and S6; Table S2). The latter two animals showed intestinal inflammation comparable to that in mice that received CL4-CD8 T cells and *S.* Tm^{avir} (data not shown). At the phylum level, we did not detect any significant differences between the microbiota recovered from the feces of the unmanipulated mice (no gut inflammation), the VILLIN-HA transgenic mice that had received CL4-CD8 T cells (gut inflammation), and the *S.* Tm^{avir}–infected VILLIN-HA transgenic mice that had not received CL4-CD8 T cells (no gut inflammation). These

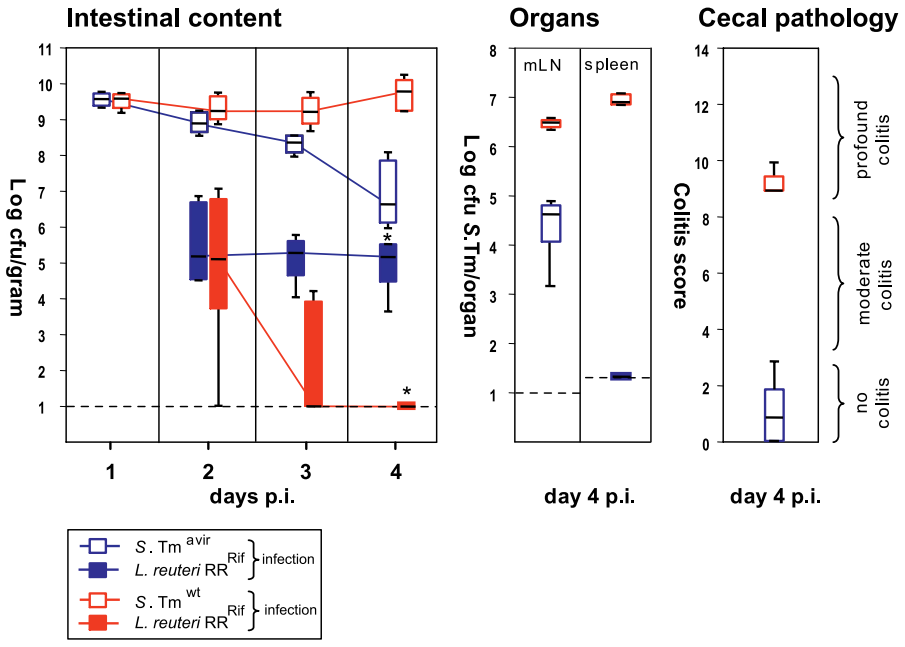


Figure 3. *S.* Tm^{wt} Can Suppress Colonization with *L. reuteri* RR^{Rif}

Groups of streptomycin-treated mice ($n = 5$) were first infected with *S.* Tm^{avir} or *S.* Tm^{wt} (5×10^7 CFU i.g.) and inoculated 1 d later with *L. reuteri* RR^{Rif} (8×10^6 CFU i.g.). Colonization levels were monitored in the feces (2 and 3 d p.i.), the cecal content (4 d p.i.), the mLN, and the spleen. Box plots show *S.* Tm^{avir} (open blue boxes), *S.* Tm^{wt} (open red boxes), *L. reuteri* RR^{Rif} in *S.* Tm^{avir}–infected mice (filled blue boxes), and *L. reuteri* RR^{Rif} in *S.* Tm^{wt}–infected mice (filled red boxes). In all groups cecal pathology was scored at day 4 p.i. *, $p \leq 0.05$; statistically significant difference in *L. reuteri* RR^{Rif} colonization between *S.* Tm^{avir}– and *S.* Tm^{wt}–infected mice. *L. reuteri* RR^{Rif} was not detected in mLN and spleen. Boxes indicate 25th and 75th percentiles, black bars indicate medians, and whiskers indicate data ranges.
doi:10.1371/journal.pbio.0050244.g003

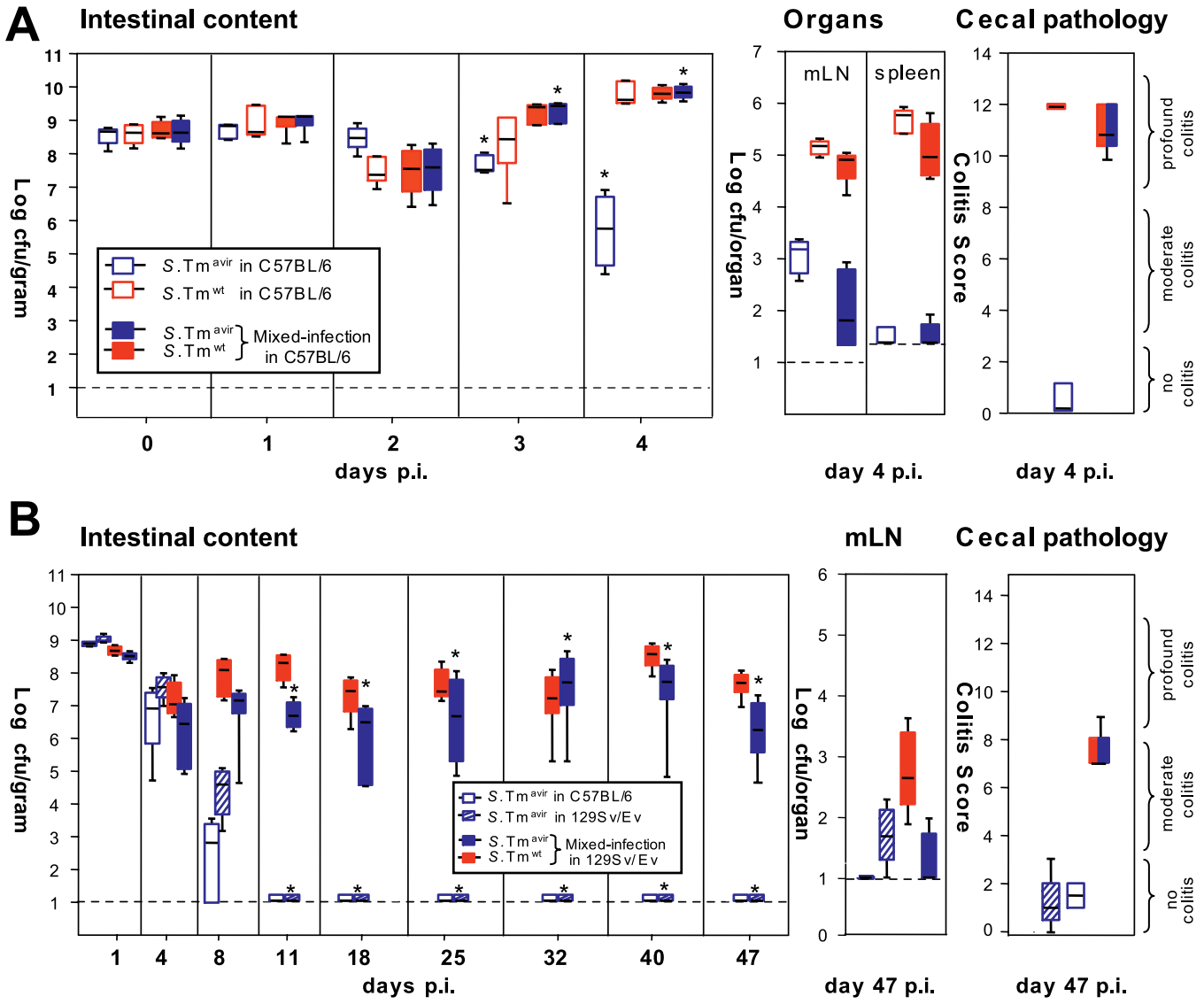


Figure 4. *S.* Tm^{wt}–Induced Inflammation Enhances Colonization of *S.* Tm^{avir}

(A) Mixed infection with *S.* Tm^{wt} complements the colonization defect of *S.* Tm^{avir}. Streptomycin-treated C57BL/6 mice ($n = 5$/group) were infected with 5×10^7 CFU i.g. of *S.* Tm^{avir} only (open blue boxes), *S.* Tm^{wt} only (open red boxes), or a 1:1 mixture of the two strains (filled blue and red boxes, respectively). Colonization was measured in the feces (days 0–3 p.i.) and the cecal content (day 4 p.i.) (left panel). Colonization of mLN and spleen (middle panel) and cecal pathology (right panel) were determined at day 4 p.i.

(B) Mixed infection with *S.* Tm^{wt} complements the colonization defect of *S.* Tm^{avir} in a chronic *Salmonella* colitis model (129Sv/Ev mice). Groups of streptomycin-treated mice (NRAMP$^+$ 129Sv/Ev mice raised by C57BL/6 foster mice; $n = 4$ per group) were infected with 5×10^7 CFU i.g. of *S.* Tm^{avir} only (blue-striped boxes) or a 1:1 mixture of *S.* Tm^{avir} and *S.* Tm^{wt} (filled blue and red boxes, respectively). One additional control group (four streptomycin-treated C57BL/6 mice) was infected with *S.* Tm^{avir} (5×10^7 CFU i.g.; open blue boxes). Colonization was measured in the feces (days 1–40 p.i.) and the cecal content (day 47 p.i.) (left panel). Colonization of mLN and spleen (middle panel) and cecal pathology (right panel) were analyzed at day 47 p.i. Boxes indicate 25th and 75th percentiles, black bars indicate medians, and whiskers indicate data ranges.

doi:10.1371/journal.pbio.0050244.g004

data suggest that inflammation per se does not drastically alter the gross gut flora composition (at least not in the short term). Further work is required to determine whether the loss of colonization resistance in the inflamed VILLIN-HA transgenic mice is attributable to suppression of some particular, low abundance member(s) of the microbiota.

Finally, our data show that *S.* Tm^{avir} colonization efficiency in the murine intestine is restricted by the intestinal microbiota. In the absence of microbiota, *S.* Tm^{avir} should colonize efficiently. This was confirmed in germ-free mice that lack microbiota in the first place. *S.* Tm^{avir} colonized the large intestine of germ-free mice at wild-type levels up to day 4 p.i. (approximately 10^9 CFU/g) but did not cause colitis (Figure S5). Thus, *S.* Tm^{avir} efficiently colonizes the murine intestine as long as competing microbiota is lacking. Furthermore, inflammation is not required for colonizing the intestinal lumen in the absence of microbiota. However, it should be noted that germ-free mice represent a useful but highly artificial tool. In natural habitats, *Salmonella* spp. always encounters a dense intestinal microbiota, and intestinal

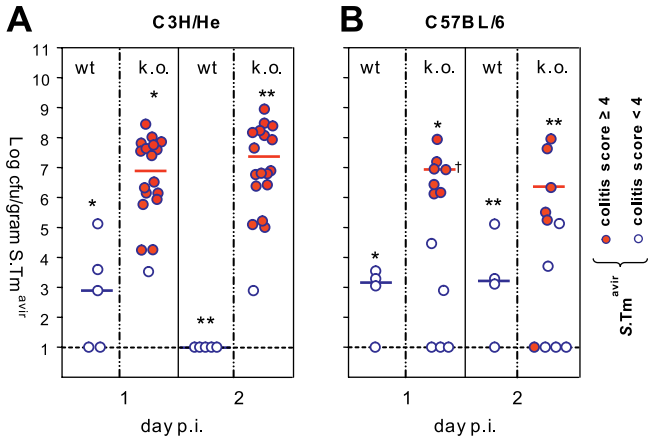


Figure 5. Intestinal Inflammation in IL10$^{-/-}$ Mice Enhances Colonization of *S.* Tm^{avir}

(A) C3H/HeJBir$^{IL10-/-}$ ($n = 18$) and C3H/HeJ control animals ($n = 5$) were infected with *S.* Tm^{avir} (5×10^7 CFU i.g.; no streptomycin treatment). *S.* Tm^{avir} colonization was analyzed in feces (day 1 p.i.) and cecum content (day 2 p.i.), and cecal pathology was scored (see Material and Methods). Open blue circles indicate mice with colitis score < 4; blue circles with red filling indicate mice with colitis score ≥ 4. *, $p = 0.03$; **, $p = 0.004$. (B) C57BL/6$^{IL10-/-}$ ($n = 12$) and C57BL/6 control animals ($n = 4$) were infected with *S.* Tm^{avir} (5×10^7 CFU i.g.; no streptomycin treatment). *S.* Tm^{avir} colonization and cecal pathology were analyzed as described above. Open blue circles indicate mice with colitis score < 4; blue circles with red filling indicate mice with colitis score ≥ 4. *, $p = 0.006$; **, $p = 0.016$. †One animal was sacrificed at the end of day 1 p.i. for humane reasons.

doi:10.1371/journal.pbio.0050244.g005

colonization will be enhanced by the triggering of inflammation.

Discussion

Based on these data we propose a three-way microbiota–pathogen–host interaction model for murine *Salmonella* colitis (Figure 7). The resident microbiota and the incoming pathogen compete for growth. In a "healthy" intestine the normal microflora is shaped and stabilized by mutually beneficial interactions with the intestinal mucosa. It effectively excludes *S.* Tm^{wt} and *S.* Tm^{avir} from the intestinal lumen. Colonization resistance can be transiently alleviated by streptomycin treatment. Inflammatory host responses—triggered by specific *S.* Tm virulence factors (TTSS-1 and TTSS-2), by genetic pre-disposition (IL10$^{-/-}$), or by T cell–inflicted damage (VILLIN-HA$^{CL4\text{-}CD8}$ model)—alter conditions in the intestinal lumen and shift the competition in favour of the incoming pathogen. Suppression of the microbiota or enhanced pathogen growth may be involved (Figure 7). In either case, *S.* Tm^{wt} can enhance intestinal colonization via an indirect mechanism—by triggering the host's immune defence. Thus, *S.* Tm^{wt} infection involves two different steps: triggering inflammation, and surviving in and profiting from the altered ecological niche. The avirulent mutant *S.* Tm^{avir} is unable to trigger colitis but it is still capable of taking advantage of the ecological niche opened by inflammation and thus successfully competes with the microbiota if inflammation is induced by other means.

How does intestinal inflammation subvert colonization resistance? The inflammation involves increased secretion of antibacterial peptides and lectins [21,22] and mucins (B. Stecher and W. Hardt, unpublished data), phagocyte infiltration/transmigration, and release of oxygen and nitrogen radicals. Potentially, there are a number of different ways this may subvert colonization resistance. (1) Released antibacterial factors may kill or retard growth of specific members of the microbiota that would normally inhibit *S.* Tm growth in the healthy intestine. (2) There may be "commensal network disruption", i.e., loss of one or more specific microbiota species that might be required for efficient growth of the microbiota species that slow pathogen growth in the normal, healthy intestine. These protecting species and their integration into microbiota growth networks have not been identified. (3) There may be differential defence susceptibility. Microbiota species conferring colonization resistance might be susceptible to antibacterial defences that *S.* Tm can resist. This would be in line with the discovery of numerous *S.* Tm genes that function to enhance antimicrobial peptide resistance and radical detoxification [23–25]. (4) There may be enhanced pathogen growth. The altered nutrient mix available in the inflamed gut might foster efficient pathogen replication. Under these conditions, microbiota may simply grow slower and are thus overgrown by the pathogen. The model is summarized in Figure 7. Future work will have to address which of these mechanisms contribute to subversion of gut inflammation by *S.* Tm.

Inflammation induced by *S.* Tm, self-reactive T cells, or IL-10 deficiency enhances colonization by the pathogen and reduces growth of the commensal microbiota. Other proteobacteria closely related to *S.* Tm may also benefit from inflammation (e.g., *E. coli;* see Figure 2). Thus, this principle may also apply to other enteric infections. For example, in calves, which are naturally susceptible to *Salmonella* enterocolitis, defects of *Salmonella* TTSS-2 mutants in triggering inflammation are associated with attenuation of intestinal colonization [26,27]. Similar observations were made with *Shigella flexneri, Vibrio cholerae,* and *Citrobacter rodentium,* the causative agents of bacillary dysentery, cholera, and transmissible murine colonic hyperplasia: ablation of colitis by disrupting the hosts' innate immune response or specific bacterial virulence factors coincided with reduced intestinal colonization [28–31]. Thus, intestinal inflammation and efficient colonization may be linked in a broad range of enteropathogenic infections.

Some data are available for human *Salmonella* enterocolitis. In line with findings in the murine system, antibiotics are known to reduce human colonization resistance, and altered microbiota composition is commonly observed in patients with inflammatory bowel disease (IBD) [32–34]. Furthermore, some studies suggest an increased incidence of *Salmonella* colonization in IBD patients [35–40].

Microbiota composition in IBD patients significantly differs from that in healthy controls. Currently, an imbalance in normal gut microbiota is regarded as one possible factor triggering the inflammation in Crohn disease and ulcerative colitis [41–43]. Our data suggest that the altered gut flora might not be the cause, but rather one of the many symptoms, of intestinal inflammation in IBD patients. Further investigation into this idea will be of importance for basic research exploring the aetiology and pathogenesis of Crohn disease and ulcerative colitis.

The outcome of any infection is determined through competition between the bacterial virulence factors (enhanc-

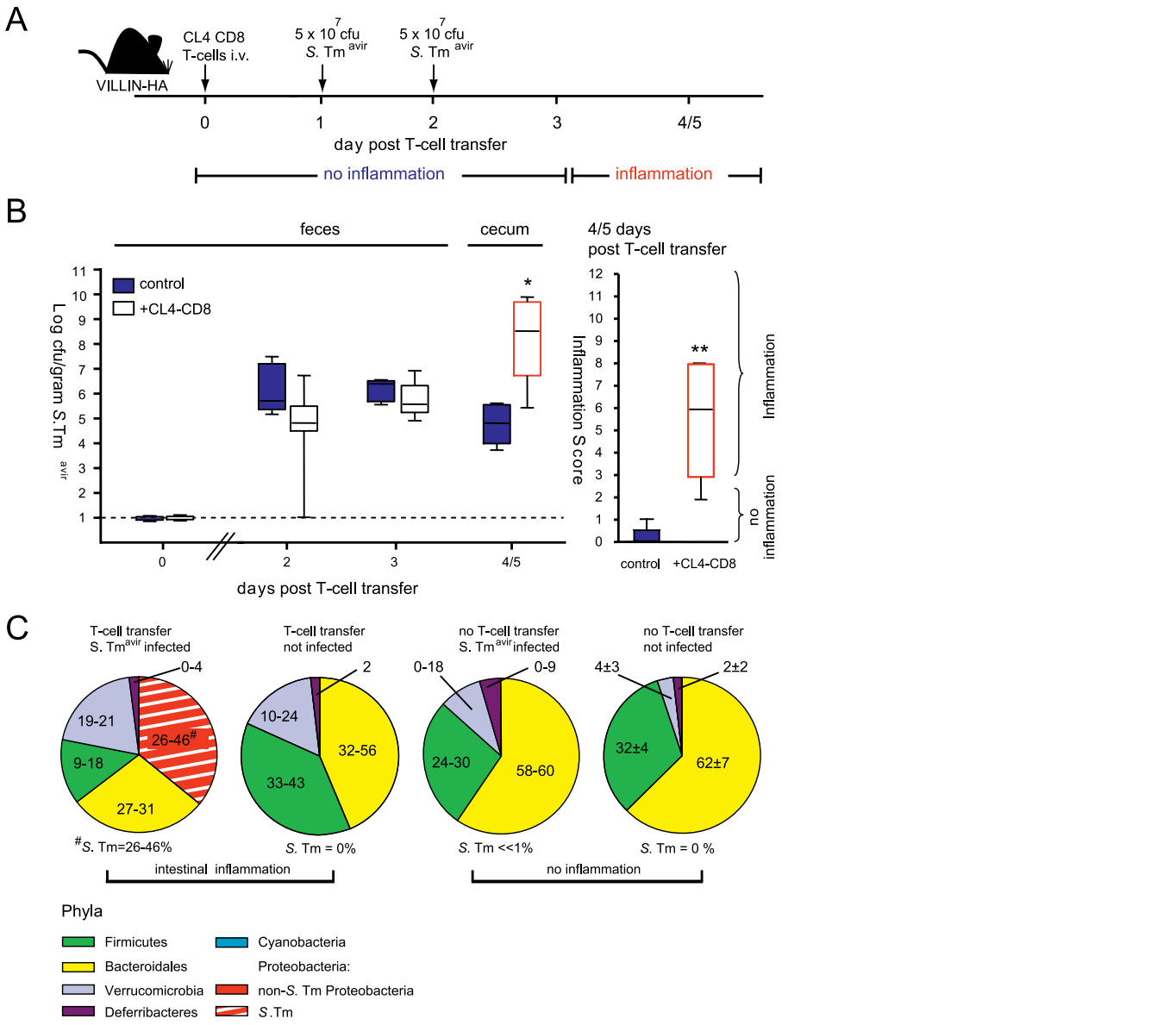


Figure 6. Gut Inflammation in the VILLIN-HA$^{CL4\text{-}CD8}$ Model Boosts *S.* Tm^{avir} Colonization

(A) The VILLIN-HA$^{CL4\text{-}CD8}$ model including the time course of intestinal inflammation and the infection regime employed in the experiment shown below.

(B) Gut colonization by *S.* Tm^{avir} is enhanced when inflammation occurs. Seven VILLIN-HA mice received 4×10^6 CL4-CD8 T cells (open white boxes) at day 0. Five unmanipulated VILLIN-HA transgenic mice served as control (blue boxes; no T cell transfer). Both groups of mice were inoculated with 5×10^7 CFU i.g. of *S.* Tm^{avir} at days 1 and 2. *S.* Tm^{avir} colonization was measured in the feces (days 1 and 2 p.i.). When symptoms of colitis (weight loss and diarrhoea) were observable in the animals from the experimental group (day 4/5), mice were sacrificed and *S.* Tm^{avir} loads in the cecal content (left) as well as cecal pathology (right) were determined (open red boxes indicate inflammation). *, $p = 0.01$; **, $p = 0.003$. Boxes indicate 25th and 75th percentiles, black bars indicate medians, and whiskers indicate data ranges.

(C) Pie diagrams showing the fecal microbiota composition at the phylum level. The average for $n = 2$ animals per group (approximately 100 16S rRNA gene sequences per animal) is shown for all groups except "no T cell transfer, not infected", for which the average for four mice is shown. Information at higher taxonomic resolution is provided in Table S1. The *p*-values are shown in Table S2.

doi:10.1371/journal.pbio.0050244.g006

ing pathogen replication/persistence) and the host's immune defences (eliminating the pathogen). In the case of enteropathogens, which target a niche colonized by the microbiota, the virulence factors can serve an additional function that has remained unrecognized: they allow triggering of intestinal inflammation that subverts the host's immune defences for undermining colonization resistance. This may represent a common virulence strategy of enteropathogenic bacteria including *Clostridium difficile*, which is a frequent cause of antibiotic-associated colitis. In fact, inflammation may promote pathogen competitiveness at any colonized site of the human body, and pathogens infecting the respiratory tract, the uro-genital system, and the skin might also use this strategy. Molecular analysis of the complex three-way pathogen–host–microbiota interactions poses a great technological challenge for future research and promises to reveal

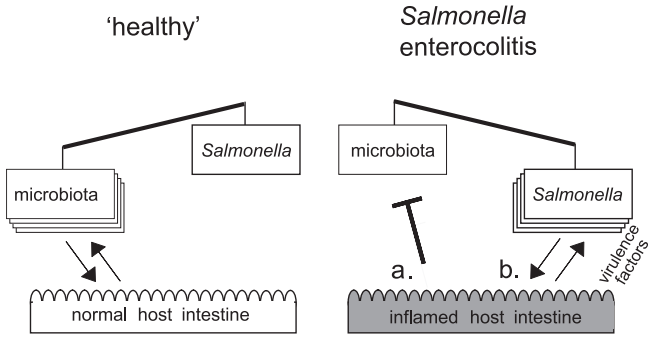


Figure 7. Working Model for the Microbiota–Host–Pathogen Interaction in Health and Disease

Colonization resistance (or lack thereof) results from growth competition between microbiota and incoming pathogens. Host responses can skew growth conditions in the intestinal lumen in either direction. Left: the normal microbiota is shaped by mutually beneficial interactions with the intestinal mucosa and mediates colonization resistance against incoming pathogens. Right: *S.* Tm employs specific virulence factors for triggering colitis. Inflammation alters the luminal conditions and shifts the growth competition in favour of the pathogen, thus alleviating colonization resistance. Inhibitory effects on the microbiota (a) and/or improved growth conditions for the pathogen (b) may be involved. Furthermore, the microbiota–pathogen growth competition can be affected by antibiotic treatment or by pre-existing intestinal inflammation.
doi:10.1371/journal.pbio.0050244.g007

novel avenues for determining prevention strategies and cures for infectious disease.

Materials and Methods

Animals. All aspects of animal procedures were approved by local authorities and performed according to the legal requirements. Sex- and age-matched specified pathogen free (SPF) C57Bl/6 (Elévage Janvier; http://www.janvier-breedingcenter.com/), 129Sv/Ev, C3H/He (Charles River Laboratories, http://www.criver.com/), C57BL/6$^{IL10-/-}$ [19], and C3H/HeJBir$^{IL10-/-}$ [18] mice were held under barrier conditions at the Rodent Centre, Swiss Institute of Technology Zurich, Zurich, Switzerland, and the Biologisches Zentrallabor, University of Zurich, Zurich, Switzerland. VILLIN-HA [44] and CL4-TCR [45] transgenic mice were raised under SPF barrier conditions at the Helmholtz Centre for Infection Research, Braunschweig, Germany, and transferred to the Rodent Centre 1 wk before the infection experiment. Germ-free C57BL/6 mice were bred and infected in the germ-free facility of the Biologisches Zentrallabor. 129Sv/Ev mice used for long-term infection experiments (Figure 4) were transferred to C57BL/6 foster mice at the day of birth, and raised and weaned as usual.

In the streptomycin mouse model, mice were treated with streptomycin (20 mg i.g.) [13] and infected 24 h later with *S.* Tm strains (5×10^7 CFU i.g.) as indicated. For super-infection, *L. reuteri* RR^{Rif} (8×10^6 CFU i.g.) was administered 24 h after *S.* Tm infection. No streptomycin treatment was performed in spontaneous colitis models and germ-free mice (Figure 3C and 3D).

For induction of acute colitis, CD8$^+$ T cells from CL4-TCR transgenic mice that express an α/βT cell receptor recognizing an epitope of the HA protein presented by MHC class I (the H-2Kd:HA512–520 complex) were adoptively transferred into VILLIN-HA mice that express the A/PR8/34 HA epitope from influenza virus A under control of the enterocyte-specific villin promoter [20]. Single-cell suspensions were prepared from the spleen of CL4-TCR transgenic mice. Cell suspensions were depleted of CD4$^+$, CD11b$^+$, CD45R$^+$, DX5$^+$, and Ter-119$^+$ cells by using the MACS CD8 T cell isolation kit (Miltenyi Biotec, http://www.miltenyibiotec.com/). CL4-TCR T cells were purified by negative selection according to the manufacturer's instructions. Isolated CD8$^+$ T cells were washed once in PBS and resuspended (4×10^7 cells/ml of PBS). Then 4×10^6 purified CL4-TCR transgenic T cells were injected intravenously into VILLIN-HA transgenic mice. Disease symptoms (weight loss and diarrhoea) were observed 4–5 d after adoptive transfer.

Bacteria. The streptomycin-resistant wild-type strain *S.* Tm^{wt} (SL1344 wild-type [46]) and the isogenic mutant *S.* Tm^{avir} (Δ*invG sseD::aphT; kan*^{R} [13]) were grown in LB 0.3 M NaCl as described [13]. Colonization was defined by plating on MacConkey agar plates (Oxoid, http://www.oxoid.com/; 100 μg/ml streptomycin). Co-infections with *S.* Tm^{avir} were evaluated by replica-plating on medium containing kanamycin (50 μg/ml).

Culturable intestinal microbiota were grown on Wilkins Chalgren agar supplemented with 5% defibrillated sheep blood (Oxoid) for 3–5 d in an atmosphere of 7% H_2, 10% CO_2, and 83% N_2 at 37 °C in anaerobic jars. 16S rRNA gene sequencing was performed as described below. *L. reuteri* RR^{Rif} was selected on MRS medium (100 μg/ml of rifampicin; Laboratoire Labo'Life, http://www.labolife.com/) and grown anaerobically.

Analysis of bacterial loads in intestinal content and systemic organs. Fresh fecal pellets collected from individual mice and cecum content were resuspended in PBS. Mesenteric lymph nodes (mLN), spleen, and liver were removed aseptically and homogenized in cold PBS (0.5% tergitol and 0.5% BSA). Bacteria were enumerated by plating on appropriate medium.

Bacterial 16S rRNA gene amplification. Colonies were isolated and purified twice on Wilkins Chalgren agar (5% sheep blood). DNA was recovered by lysis (Tris/EDTA; 0.5% SDS and 0.1 mg/ml of proteinase K; 37 °C; 1 h), CTAB treatment (1%; 62.5 mM NaCl; 65 °C; 10 min), phenol-chloroform extraction, and 2-propanol precipitation. Broad-range bacterial primers fD_1 (5′-AGA GTT TGA TCC TGG CTC AG-3′) and rP_1 (5′-ACG GTT ACC TTG TTA CGA CTT-3′) [47] were used for 16S rRNA gene PCR amplification (94 °C, 5 min; 35 cycles of 94 °C, 1 min; 43 °C, 1 min; 72 °C, 2 min; and 7-min final extension at 72 °C). The PCR product was purified and sequenced with primer rP_1.

Quantification of cultured bacteria. First, bacteria were grouped according to colony morphology. Then, representative colonies were typed by 16S rRNA gene sequencing and comparison to the Ribosomal Database Project II [48]. This allowed a rough estimation of the abundance of the respective bacterial species (Table 1). Two mice were analyzed per condition (*S.* Tm^{wt} and *S.* Tm^{avir} infection day 4 p.i.). Six colony morphological groups were assigned for *S.* Tm^{wt} infection, and ten for *S.* Tm^{avir} infection.

Histopathological evaluation. Tissues were cryo-embedded in Tissue Tek OCT Compound (Sysmex, http://www.sysmex-europe.com/), 5-μm cryosections were stained with hematoxylin and eosin (HE), and cecum pathology was evaluated using a histopathological scoring scheme as previously described [49,50] (see Figure S1). Evaluation scored submucosal edema (score 0–3), polymorphonuclear leukocyte infiltration into the lamina propria (score 0–4), loss of goblet cells (score 0–3), and epithelial damage (score 0–3). The combined pathological score for each tissue sample was determined as the sum of these averaged scores: 0–3, no to minimal signs of inflammation that are not sign of a disease (this is frequently found in the cecum of SPF mice); 4–8, moderate inflammation; and 9–13, profound inflammation.

Immunofluorescence microscopy. Cecal tissues were fixed in PBS (4% paraformaldehyde [pH 7.4]; 4 °C; 12 h), washed in PBS, equilibrated in PBS (20% sucrose and 0.02% NaN_3; 4 °C; 12 h) and cryo-embedded in OCT. Cryosections (7 μm) were mounted on glass slides, air-dried (21 °C; 2 h), fixed in PBS (4% paraformaldehyde, 5 min), washed, and blocked with 10% (w/v) normal goat serum in PBS (1 h). *S.* Tm was stained with polyclonal rabbit anti–*Salmonella* O antigen group B serum (factors 1, 4, 5, and 12, Brunschwig, http://www.brunschwig-ch.com/; 1:500 in PBS, 10% goat serum) and a Cy3-conjugated goat anti-rabbit antibody (Milan; 1:300 in PBS, 10% goat serum). The specificity of the anti–*Samonella* O (1, 4 ,5, and 12) antiserum was checked extensively by immunofluorescence microscopy. This was done by analyzing cecum tissue sections from uninfected mice (negative), *S.* Tm–infected mice (positive), *S. enterica* serovar Enteritidis–infected mice (negative; the LPS of this serovar does not react with this antiserum), and mice with >10 different commensal species, including commensal *E. coli* strains from our mouse colony, grown in vitro (all negative). DNA was stained with Sytox green (0.1 μg/ml; Sigma-Aldrich, http://www.sigmaaldrich.com/) and F-Actin with Alexa-647-phalloidin (Molecular Probes, http://probes.invitrogen.com/). Sections were mounted with Vectashield hard set (Vector Laboratories, http://www.vectorlabs.com/) and sealed with nail polish. Images were recorded using a PerkinElmer (http://www.perkinelmer.com/) Ultraview confocal imaging system and a Zeiss (http://www.zeiss.com/) Axiovert 200 microscope. For quantification of total bacterial numbers, cecal contents were weighed, fixed in 4% paraformaldehyde, and stained with Sytox green (0.1 μg/ml). Bacteria were counted in a Neubauer's counting chamber using an upright fluorescence microscope (Zeiss).

Broad-range bacterial 16S rRNA gene sequence analysis. Total DNA was extracted from cecal contents using a QIAmp DNA stool mini kit (Qiagen, http://www1.qiagen.com/) and a Tissuelyzer device

(Qiagen). 16S rRNA genes were amplified by PCR using primers Bact-7F (5′-AGA GTT TGA TYM TGG CTC AG-3′) and Bact-1510R (5′-ACG GYT ACC TTG TTA CGA CTT-3′) and the following cycling conditions: 95 °C, 5 min; 22 cycles of 95 °C, 30 s; 58 °C, 30 s; 72 °C, 2 min; followed by 72 °C, 8 min; 4 °C, ∞. Reaction conditions (100 μl) were as follows: 50 mM KCl, 10 mM Tris-HCl (pH 8.3), 1.5 mM Mg^{2+}, 0.2 mM dNTPs, 40 pmol of each primer, and 5 U of Taq DNA polymerase (Eppendorf, http://www.eppendorf.com/). Fragments were purified by gel electrophoresis, excised, recovered using the gene clean kit (Qbiogene; http://www.qbiogene.com/) and dried. The PCR products were suspended in 10 μl of sterile distilled water and between 2 and 5 μl was ligated into pGEM-T Easy Vectors (Promega, http://www.promega.com/). The ligated vectors were transformed into high-efficiency competent JM109 *E. coli* cells (Promega), plated on LB-carbenicillin agar, and subjected to blue-white screening of colonies. White colonies were picked into 96-well boxes containing 500 μl of Circlegrow medium (Qbiogene, http://www.qbiogene.com/) per well and grown overnight at 37 °C, and the plasmid DNA was then prepped using a modified semi-automated alkaline lysis method. Sequencing was carried out using Applied Biosystems (http://www.appliedbiosystems.com/) BigDye terminators (version 3.1) and run on Applied Biosystems 3730 sequencers. The 16S rRNA gene inserts were sequenced using two primers targeted towards the vector end sequences, M13r (5′-CAGGAAACAGCTATGACC-3′) and T7f (5′-TAATACGACTCACTATAGGG-3′), and one towards an internal region of the gene, 926r (5′-CCGTCAATTC[A/C]TTT[A/G]AGT-3′), in order to bridge any gaps between the sequences generated from the two end primers.

Contigs were built from each three-primer set of sequences using the GAP4 software package [51] and converted to "sense" orientation using OrientationChecker software [52]. These files were then aligned using MUSCLE [53], and the alignments were manually inspected and corrected using the sequence editor function in the ARB package [54]. The files were then tested for the presence of chimeric sequences using Mallard [52] and Bellerophon [55], and putative chimeras were checked using Pintail [56] and BLAST [57]. Positively identified chimeras were removed, and the remaining sequences were examined with the Classifier function at the Ribosomal Database Project II Web site [48] in order to give a broad classification at the phylum level. To obtain more detailed taxonomic information the sequences were divided into phylotypes by generating distance matrices in ARB (with Olsen correction), which were then entered into the DOTUR program [58] set to the furthest neighbour and 99% similarity settings. The resulting phylotypes were then assigned similarities to nearest neighbours using BLAST.

Statistical analysis of bacterial colonization and intestinal pathology. Statistical analyses of viable CFU and pathological scores were performed using the exact Mann-Whitney *U* Test and the SPSS version 14.0 software, as described before [8]. Values of $p < 0.05$ were considered statistically significant. Box-plots were created using GraphPad Prism 4 version 4.03 (GraphPad Software, http://www.graphpad.com/).

Statistical analysis of microbiota composition. Differences in the phylogenetic compositions of samples were assessed by first assigning the detected 16S rRNA gene sequences to their respective phyla, and then computing the normalized Euclidean distance between the phyla counts. The observed differences were judged for their statistical significance by performing Monte Carlo randomizations: 16S rRNA gene sequences were shuffled between two samples, such that overall sample sizes and total counts for each phylum were maintained. Euclidean distances were then re-computed, and the fraction of distances larger than or equal to the observed distances determined the *p*-values. Bonferroni correction for multiple testing means that *p*-values below 0.005 indicate statistical significance in Figures 2 and 6 and Table 2.

Supporting Information

Figure S1. Colitis Score Developed for the Streptomycin-Pretreated Mouse Model for *Salmonella* Colitis [8]

Mice were pretreated with a single dose of streptomycin (20 mg i.g.) and 24 h later infected with 5×10^7 CFU of *S.* Tm^{avir} (A) or *S.* Tm^{wt} i.g. (B). Mice were sacrificed 1 d p.i.

Left panels of (A) and (B): macroscopic appearance of the cecum from *S.* Tm^{avir}– and *S.* Tm^{wt}–infected mice, respectively. Note the reduction in size and purulent cecal content in case of *S.* Tm^{wt}–induced colitis.

Middle panels: HE-stained cross-section of ceca shown in left panel (scale bar: 1 mm). Note the submucosal edema (se), which is a characteristic of *S.* Tm^{wt}–induced colitis. L, cecal lumen.

Right panels: at higher magnification, large numbers of goblet cells (gc) are observed in the cecal mucosa of healthy mice. Colitis leads to reduced numbers of goblet cells due to pronounced epithelial regeneration. Note infiltrating polymorphonuclear leukocytes and desquamated epithelium in the *S.* Tm^{wt}–infected cecum (scale bar: 0.05 mm).

Detailed parameters for colitis score are listed in table at bottom of figure.

Found at doi:10.1371/journal.pbio.0050244.sg001 (272 KB PDF).

Figure S2. FISH Analysis of Microbiota Manipulation by *S.* Tm^{wt} and *S.* Tm^{avir} in the Streptomycin Mouse Model

Cecal contents were fixed in PBS (4% paraformaldehyde [pH 7.4]; 4 °C; 12 h), washed in PBS, applied onto polylysine-coated slides, and air-dried. Bacteria were permeabilized (70.000 U/ml of lysozyme; 5 mM EDTA; 100 mM Tris/HCl [pH 7.5]; 37 °C; 10 min), dehydrated with ethanol, and hybridized with HPLC-purified, 5′-labelled 16S rRNA probes (5% formamide, 90 mM NaCl, 20 mM Tris/HCl [pH 7.5]; 46 °C; 2 h): Eub338-cy5 (5′-GCT GCC TCC CGT AGG AGT-3′; detection of all eubacteria [59]), LGC-cy3 or LGC-fluorescein (5′-TCA CGC GGC GTT GCT C-3′; detection of gram-positive bacteria with low G+C content; Firmicutes [60]), and Bac303-cy3 or Bac303-fluorescein (5′-CCA ATG TGG GGG ACC TT-3′; detection of the Bacteroidales group of the Bacteroidetes [61]). Slides were washed at 48 °C (636 mM NaCl, 5 mM EDTA, 0.01% SDS, 20 mM Tris/HCl [pH 7.5]) as described [59]. *S.* Tm was detected by immunostaining (see above), and FISH detection was performed using the Eub338-cy5 probe. The relative abundance of Firmicutes, Bacteroidales, and *S.* Tm was determined by co-staining and imaging at 630× magnification using a PerkinElmer Ultraview confocal imaging system and a Zeiss Axiovert 200 microscope. For each condition, 500–1,750 bacteria were evaluated.

FISH analysis of cecal microbiota from the mice shown in Figure 2. Cecal contents from unmanipulated mice, from mice at days 1 or 5 after streptomycin treatment (20 mg, i.g.), and from streptomycin-treated mice 4 d after infection with *S.* Tm^{avir} and *S.* Tm^{wt} (5×10^7 CFU i.g.; all $n = 5$) were recovered, fixed on cover slips, and hybridized with Eub338 (all bacteria). Firmicutes and Bacteroidales were recognized by hybridization with LGC and BAC303 probes, respectively, and *S.* Tm by an anti–*S.* Tm LPS antiserum (see Materials and Methods). Firmicutes (green), $Eub338^+$ $Bac303^-$ LGC^+; Bacteroidales (yellow), $Eub338^+$ $Bac303^+$ LGC^-; *Salmonella* (red with white stripes), $Eub338^+$ LPS^+; "unknown" (grey), $Eub338^+$ LGC^- $Bac303^-$ LPS^-. Abundance of respective groups is expressed as percentage of total $Eub338^+$ bacteria.

The results of the FISH analysis confirmed the results obtained via 16S rRNA gene sequencing (Figure 2). Slight differences in the percent composition of the microbiota with respect to Firmicutes, Bacteroidales, and *Salmonella* spp. obtained via both methods are attributable to species-specific differences in lysis efficiency and 16S rRNA gene copy number.

Found at doi:10.1371/journal.pbio.0050244.sg002 (124 KB PDF).

Figure S3. Cecal Histopathology in Acute and Chronic Mouse Colitis Models Shown in Figures 4 and 5

Frozen sections of cecal tissues (5 μm) were stained with HE (scale bar: 200 μm). Acute *Salmonella* colitis was observed in C57BL/6 mice infected with *S.* Tm^{wt} (A) but not with *S.* Tm^{avir} (B) 4 d p.i. (compare with Figure 3A). Chronic *Salmonella* colitis was observed in 129Sv/Ev mice infected with *S.* Tm^{wt} (C) but not with *S.* Tm^{avir} (D) 47 d p.i. (compare with Figure 3B). Genetic predisposition (lack of anti-inflammatory cytokine IL10) leads to sporadic occurrence of colitis in $C57/BL6^{IL10-/-}$ mice (E). However, some $C57/BL6^{IL10-/-}$ mice are not affected (F) (compare with Figure 3C). A large number of $C3H/HeJBir^{IL10-/-}$ mice were affected by cecal inflammation (G), but one was not (H) (compare with Figure 3C). L, cecal lumen; se, submucosal edema.

Found at doi:10.1371/journal.pbio.0050244.sg003 (735 KB PDF).

Figure S4. Colitis Scores for $C57/BL6^{IL10-/-}$ and $C3H/HeJBir^{IL10-/-}$ Mice

(A) Frozen sections of cecal tissues (5 μm) were stained with HE (scale bar: 200 μm). Histopathology was scored with respect to submucosal edema (black), polymorphonuclear leukocyte infiltration (grey), loss of goblet cells (dark grey), and epithelial destruction (light grey). The scoring scheme is shown in Figure S1. Scores are plotted as stacked

vertical bars. One animal was sacrificed at the end of day 1 p.i. for humane reasons (marked with †).
(B) Confocal fluorescence microscopy image of cecal lumen reveals normal high microbiota densities. Upper left: C3H/HeJBir$^{IL10-/-}$ animal marked with ‡ in (A). The remaining images show animals described in Figure 6B. Upper right: VILLIN-HA control, *S.* Tm^{avir} infected. Lower left: VILLIN-HA+CL4-CD8 (inflammation), non-infected. Lower-right: VILLIN-HA+CL4-CD8 (inflammation), *S.* Tm^{avir} infected. Bacterial DNA is stained by Sytox green (green) and extracellular *S.* Tm by anti-*S.* Tm LPS antiserum (red). Scale bar: 20 or 50 μm as specified.

Found at doi:10.1371/journal.pbio.0050244.sg004 (1.8 MB PDF).

Figure S5. *S.* Tm^{avir} Efficiently Colonizes Germ-Free Mice

Germ-free C57BL/6 mice ($n = 8$) were infected with *S.* Tm^{avir} (5×10^7 CFU i.g.) and sacrificed at day 2 or 4 p.i. (open blue boxes). For comparison, previous data [62] from five mice infected for 1 d with *S.* Tm^{wt} are included (open red boxes). *S.* Tm colonization was analyzed in the cecum content (day 2 p.i.), and cecum pathology was scored (see Material and Methods). Detection limits (dotted line): cecum, 10 CFU/g; mLN, 10 CFU/organ; spleen, 20 CFU/organ. At day 4 p.i., *S.* Tm^{avir} colonization levels in germ-free mice in the absence of re-growing microbiota were significantly higher when compared to streptomycin-treated SPF mice ($p = 0.002$; compare with Figure 3A, left panel).

Found at doi:10.1371/journal.pbio.0050244.sg005 (105 KB PDF).

Figure S6. 16S rRNA Gene Sequence Analysis of Microbiota in VILLIN-HA$^{CL4-CD8}$ Model

Visual depiction of the microbiota composition of individual mice. The animals were grouped based on the similarity of their microbiota composition at the phylum level (using the Canberra distance as metric). The resulting groupings are depicted as a dendrogram, and observed phylum counts for each mouse are shown as a heat map (0%–100% of all identified 16S rRNA gene sequences). Labels give unique mouse identifier numbers. The experimental groups are indicated.

Found at doi:10.1371/journal.pbio.0050244.sg006 (64 KB PDF).

Table S1. Broad-Range Bacterial 16S rRNA Gene Sequence Analysis of the Microbiota Composition from the Experiment Shown in Figures 2 and 6

Found at doi:10.1371/journal.pbio.0050244.st001 (277 KB XLS).

Table S2. Phylum-Level Comparison of Microbiota of VILLIN-HA$^{CL4-CD8}$ Model from the Experiment Described in Figure 6

Found at doi:10.1371/journal.pbio.0050244.st002 (35 KB DOC).

Accession Numbers

The GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) accession numbers for the 16S RNA gene sequences shown in Figure 2 are EF604903–EF605247, and for those shown in Figure 6C are EF604904–EF605247 and EU006095–EU006496.

Acknowledgments

The authors are grateful to Paul Scott for construction of the clone libraries and Carol Churcher and the Pathogen Sequencing Unit at the Sanger Institute for sequencing. We thank Siegfried Hapfelmeier and Mathias Heikenwälder for discussion, C. Sigurdson for C57BL/6$^{IL10-/-}$ mice, and Ryan Mcarthur for reading the manuscript.

Author contributions. BS, RR, AWW, AMW, JB, JP, GD, CvM, and WDH conceived and designed the experiments. BS, RR, AWW, AMW, MB, and AJM performed the experiments; AWW performed microbiota analysis by 16S gene sequencing. BS, RR, AWW, MK, SC, CvM, and WDH analyzed the data. BS and WDH wrote the paper.

Funding. This work was supported by a grant to WDH from the Swiss National Science Foundation (#3100A0–100175/1) and by the Wellcome Trust (AWW, GD, and JP).

Competing interests. The authors have declared that no competing interests exist.

References

1. Ley RE, Peterson DA, Gordon JI (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell 124: 837–848.
2. Suzuki K, Meek B, Doi Y, Muramatsu M, Chiba T, et al. (2004) Aberrant expansion of segmented filamentous bacteria in IgA-deficient gut. Proc Natl Acad Sci U S A 101: 1981–1986.
3. Sonnenburg JL, Xu J, Leip DD, Chen CH, Westover BP, et al. (2005) Glycan foraging in vivo by an intestine-adapted bacterial symbiont. Science 307: 1955–1959.
4. Samuel BS, Gordon JI (2006) A humanized gnotobiotic mouse model of host-archaeal-bacterial mutualism. Proc Natl Acad Sci U S A 103: 10011–10016.
5. Hooper LV, Gordon JI (2001) Commensal host-bacterial relationships in the gut. Science 292: 1115–1118.
6. Konrad A, Cong Y, Duck W, Borlaza R, Elson CO (2006) Tight mucosal compartmentation of the murine immune response to antigens of the enteric microbiota. Gastroenterology 130: 2050–2059.
7. Hooper LV (2004) Bacterial contributions to mammalian gut development. Trends Microbiol 12: 129–134.
8. Barthel M, Hapfelmeier S, Quintanilla-Martinez L, Kremer M, Rohde M, et al. (2003) Pretreatment of mice with streptomycin provides a *Salmonella enterica* serovar Typhimurium colitis model that allows analysis of both pathogen and host. Infect Immun 71: 2839–2858.
9. Bohnhoff M, Drake BL, Miller CP (1954) Effect of streptomycin on susceptibility of intestinal tract to experimental *Salmonella* infection. Proc Soc Exp Biol Med 86: 132–137.
10. Hapfelmeier S, Hardt WD (2005) A mouse model for *S. typhimurium*-induced enterocolitis. Trends Microbiol 13: 497–503.
11. Coburn B, Li Y, Owen D, Vallance BA, Finlay BB (2005) *Salmonella enterica* serovar Typhimurium pathogenicity island 2 is necessary for complete virulence in a mouse model of infectious enterocolitis. Infect Immun 73: 3219–3227.
12. Stecher B, Paesold G, Barthel M, Kremer M, Jantsch J, et al. (2006) Chronic *Salmonella enterica* serovar Typhimurium-induced colitis and cholangitis in streptomycin-pretreated Nramp1+/+ mice. Infect Immun 74: 5047–5057.
13. Hapfelmeier S, Ehrbar K, Stecher B, Barthel M, Kremer M, et al. (2004) Role of the *Salmonella* pathogenicity island 1 effector proteins SipA, SopB, SopE, and SopE2 in *Salmonella enterica* subspecies 1 serovar Typhimurium colitis in streptomycin-pretreated mice. Infect Immun 72: 795–809.
14. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the human intestinal microbial flora. Science 308: 1635–1638.
15. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, et al. (2005) Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A 102: 11070–11075.
16. Rawls JF, Mahowald MA, Ley RE, Gordon JI (2006) Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. Cell 127: 423–433.
17. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444: 1027–1031.
18. Lytle C, Tod TJ, Vo KT, Lee JW, Atkinson RD, et al. (2005) The peroxisome proliferator-activated receptor gamma ligand rosiglitazone delays the onset of inflammatory bowel disease in mice with interleukin 10 deficiency. Inflamm Bowel Dis 11: 231–243.
19. Kuhn R, Lohler J, Rennick D, Rajewsky K, Muller W (1993) Interleukin-10-deficient mice develop chronic enterocolitis. Cell 75: 263–274.
20. Westendorf AM, Fleissner D, Deppenmeier S, Gruber AD, Bruder D, et al. (2006) Autoimmune-mediated intestinal inflammation-impact and regulation of antigen-specific CD8+ T cells. Gastroenterology 131: 510–524.
21. Dann SM, Eckmann L (2007) Innate immune defenses in the intestinal tract. Curr Opin Gastroenterol 23: 115–120.
22. Cash HL, Whitham CV, Behrendt CL, Hooper LV (2006) Symbiotic bacteria direct expression of an intestinal bactericidal lectin. Science 313: 1126–1130.
23. Bader MW, Sanowar S, Daley ME, Schneider AR, Cho U, et al. (2005) Recognition of antimicrobial peptides by a bacterial sensor kinase. Cell 122: 461–472.
24. Navarre WW, Halsey TA, Walthers D, Frye J, McClelland M, et al. (2005) Co-regulation of *Salmonella enterica* genes required for virulence and resistance to antimicrobial peptides by SlyA and PhoP/PhoQ. Mol Microbiol 56: 492–508.
25. Uzzau S, Bossi L, Figueroa-Bossi N (2002) Differential accumulation of *Salmonella*[Cu, Zn] superoxide dismutases SodCI and SodCII in intracellular bacteria: Correlation with their relative contribution to pathogenicity. Mol Microbiol 46: 147–156.
26. Bispham J, Tripathi BN, Watson PR, Wallis TS (2001) *Salmonella* pathogenicity island 2 influences both systemic salmonellosis and *Salmonella*-induced enteritis in calves. Infect Immun 69: 367–377.
27. Coombes BK, Coburn BA, Potter AA, Gomis S, Mirakhur K, et al. (2005) Analysis of the contribution of *Salmonella* pathogenicity islands 1 and 2 to enteric disease progression using a novel bovine ileal loop model and a murine model of infectious enterocolitis. Infect Immun 73: 7161–7169.
28. Deng W, Vallance BA, Li Y, Puente JL, Finlay BB (2003) *Citrobacter rodentium* translocated intimin receptor (Tir) is an essential virulence factor needed

for actin condensation, intestinal colonization and colonic hyperplasia in mice. Mol Microbiol 48: 95–115.
29. Khan MA, Ma C, Knodler LA, Valdez Y, Rosenberger CM, et al. (2006) Toll-like receptor 4 contributes to colitis development but not to host defense during *Citrobacter rodentium* infection in mice. Infect Immun 74: 2522–2536.
30. Rabbani GH, Albert MJ, Rahman H, Islam M, Mahalanabis D, et al. (1995) Development of an improved animal model of shigellosis in the adult rabbit by colonic infection with *Shigella flexneri* 2a. Infect Immun 63: 4350–4357.
31. Sigel SP, Finkelstein RA, Parker CD (1981) Ability of an avirulent mutant of *Vibrio cholerae* to colonize in the infant mouse upper bowel. Infect Immun 32: 474–479.
32. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, et al. (2006) Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut 55: 205–211.
33. Conte MP, Schippa S, Zamboni I, Penta M, Chiarini F, et al. (2006) Gut-associated bacterial microbiota in paediatric patients with inflammatory bowel disease. Gut 55: 1760–1767.
34. Swidsinski A, Weber J, Loening-Baucke V, Hale LP, Lochs H (2005) Spatial organization and composition of the mucosal flora in patients with inflammatory bowel disease. J Clin Microbiol 43: 3380–3389.
35. Kressner MS, Williams SE, Biempica L, Das KM (1982) Salmonellosis complicating ulcerative colitis. Treatment with trimethoprim-sulfamethoxazole. JAMA 248: 584–585.
36. Hook EW (1961) Salmonellosis: Certain factors influencing the interaction of *Salmonella* and the human host. Bull N Y Acad Med 37: 499–512.
37. Taylor-Robinson S, Miles R, Whitehead A, Dickinson RJ (1989) *Salmonella* infection and ulcerative colitis. Lancet 1: 1145.
38. Lindeman RJ, Weinstein L, Levitan R, Patterson JF (1967) Ulcerative colitis and intestinal salmonellosis. Am J Med Sci 254: 855–861.
39. Isbister WH, Hubler M (1998) Inflammatory bowel disease in Saudi Arabia: Presentation and initial management. J Gastroenterol Hepatol 13: 1119–1124.
40. Szilagyi A, Gerson M, Mendelson J, Yusuf NA (1985) *Salmonella* infections complicating inflammatory bowel disease. J Clin Gastroenterol 7: 251–255.
41. Kleessen B, Kroesen AJ, Buhr HJ, Blaut M (2002) Mucosal and invading bacteria in patients with inflammatory bowel disease compared with controls. Scand J Gastroenterol 37: 1034–1041.
42. Seksik P, Rigottier-Gois L, Gramet G, Sutren M, Pochart P, et al. (2003) Alterations of the dominant faecal bacterial groups in patients with Crohn's disease of the colon. Gut 52: 237–242.
43. Gophna U, Sommerfeld K, Gophna S, Doolittle WF, Veldhuyzen van Zanten SJ (2006) Differences between Crohn's disease and ulcerative colitis patients in tissue-associated intestinal microflora. J Clin Microbiol 44: 4136–4141.
44. Westendorf AM, Templin M, Geffers R, Deppenmeier S, Gruber AD, et al. (2005) CD4+ T cell mediated intestinal immunity: Chronic inflammation versus immune regulation. Gut 54: 60–69.
45. Morgan DJ, Liblau R, Scott B, Fleck S, McDevitt HO, et al. (1996) CD8(+) T cell-mediated spontaneous diabetes in neonatal mice. J Immunol 157: 978–983.
46. Hoiseth SK, Stocker BA (1981) Aromatic-dependent *Salmonella typhimurium* are non-virulent and effective as live vaccines. Nature 291: 238–239.
47. Weisburg WG, Barns SM, Pelletier DA, Lane DJ (1991) 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol 173: 697–703.
48. Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, et al. (2005) The Ribosomal Database Project (RDP-II): Sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res 33: D294–D296.
49. Hapfelmeier S, Stecher B, Barthel M, Kremer M, Müller A, et al. (2005) The *Salmonella* pathogenicity island (SPI)-1 and SPI-2 type III secretion systems allow *Salmonella* serovar Typhimurium to trigger colitis via MyD88-dependent and MyD88-independent mechanisms. J Immunol 174: 1675–1685.
50. Stecher B, Hapfelmeier S, Muller C, Kremer M, Stallmach T, et al. (2004) Flagella and chemotaxis are required for efficient induction of *Salmonella enterica* serovar Typhimurium colitis in streptomycin-pretreated mice. Infect Immun 72: 4138–4150.
51. Staden RJD, Bonfield JK (2003) Managing sequencing projects in the GAP4 environment. In: Krawetz SA, Womble DD, editors. Introduction to bioinformatics: A theoretical and practical approach. Totawa (New Jersey): Humana Press.
52. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2006) New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. Appl Environ Microbiol 72: 5734–5741.
53. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
54. Ludwig W, Strunk O, Westram R, Richter L, Meier H, et al. (2004) ARB: A software environment for sequence data. Nucleic Acids Res 32: 1363–1371.
55. Huber T, Faulkner G, Hugenholtz P (2004) Bellerophon: A program to detect chimeric sequences in multiple sequence alignments. Bioinformatics 20: 2317–2319.
56. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl Environ Microbiol 71: 7724–7736.
57. Ye J, McGinnis S, Madden TL (2006) BLAST: Improvements for better sequence analysis. Nucleic Acids Res 34: W6–W9.
58. Schloss PD, Handelsman J (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol 71: 1501–1506.
59. Amann RI, Binder BJ, Olson RJ, Chisholm SW, Devereux R, et al. (1990) Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. Appl Environ Microbiol 56: 1919–1925.
60. Kusel K, Pinkart HC, Drake HL, Devereux R (1999) Acetogenic and sulfate-reducing bacteria inhabiting the rhizoplane and deep cortex cells of the sea grass *Halodule wrightii*. Appl Environ Microbiol 65: 5117–5123.
61. Manz W, Amann R, Ludwig W, Vancanneyt M, Schleifer KH (1996) Application of a suite of 16S rRNA-specific oligonucleotide probes designed to investigate bacteria of the phylum cytophaga-flavobacter-bacteroides in the natural environment. Microbiology 142: 1097–1106.
62. Stecher B, Macpherson AJ, Hapfelmeier S, Kremer M, Stallmach T, et al. (2005) Comparison of *Salmonella enterica* serovar Typhimurium colitis in germfree mice and mice pretreated with streptomycin. Infect Immun 73: 3228–3241.

B

TERMITES IN THE WOODWORK

Minireview

Termites in the woodwork

Samuel Chaffron and Christian von Mering

Address: Institute of Molecular Biology and Swiss Institute of Bioinformatics, University of Zurich, Winterthurerstrasse, 8057 Zurich, Switzerland.

Correspondence: Christian von Mering. Email: mering@molbio.uzh.ch

Published: 22 November 2007

Genome ***Biology*** 2007, **8:**229 (doi:10.1186/gb-2007-8-11-229)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/11/229

Abstract

Termites eat and digest wood, but how do they do it? Combining advanced genomics and proteomics techniques, researchers have now shown that microbes found in the termites' hindguts possess just the right tools.

Most animals, from insects to mammals, carry complex communities of microbes in their digestive tracts. In the case of wood-eating termites, these gut microbes are particularly important: they are thought to provide most of the capabilities needed for efficient digestion of wood, which is otherwise a largely inaccessible food source. They also help to compensate for the paucity of some nutrients in wood, for example by fixing atmospheric nitrogen, and they synthesize essential amino acids and other compounds for their hosts [1,2].

Despite their importance, relatively little is known about gut microbes in termites. This is partly because gut microbes are often difficult to grow in pure culture (as is the case for most microbes sampled from natural environments). Furthermore, a single termite can harbor a very complex assemblage of hundreds of different microbial lineages, whose members may vary widely in terms of abundance and growth rates. Without access to cultivated strains, researchers have to rely on so-called 'cultivation-independent' molecular techniques to analyze such communities. A clever combination of these techniques has now been applied to a section of the termite hindgut, aiming to identify molecular tools used by the microbes in this compartment to degrade wood [3]. Here, we review the procedures and results of this study, and discuss insights into the biological system as well as implications for the generation of biofuels.

A comprehensive inventory

As recently as 2004, biologists had rather limited experimental options for taking stock of uncultured microbes in their natural environments. They could analyze selected phylogenetic marker genes to assess taxonomic identity (using *in situ* hybridization or PCR-based sequencing), or they could use expression cloning to screen for genes encoding a specific activity of interest. Another possibility was to clone and sequence individual DNA fragments isolated from the community, in the hope of finding phylogenetic marker genes and important functional genes together on the same fragment: this latter approach can help to map lifestyles to a given lineage [4,5]. However, none of these strategies simultaneously provides a global inventory of both the taxonomic and the functional properties of a microbial community.

To overcome this limitation, researchers have since begun to apply genomics (and proteomics) technologies in high-throughput mode, analyzing entire microbial assemblages without first cloning individual strains [6-9]. These exciting new research approaches ('environmental genomics', 'metagenomics' and 'metaproteomics') put the possibility of a molecular description of an entire microbial community within reach for the first time. For the termite gut ecosystem, Warnecke and colleagues [3] have now attempted just that, in a formidable *tour de force*. They even went a step further by complementing their work with a preliminary

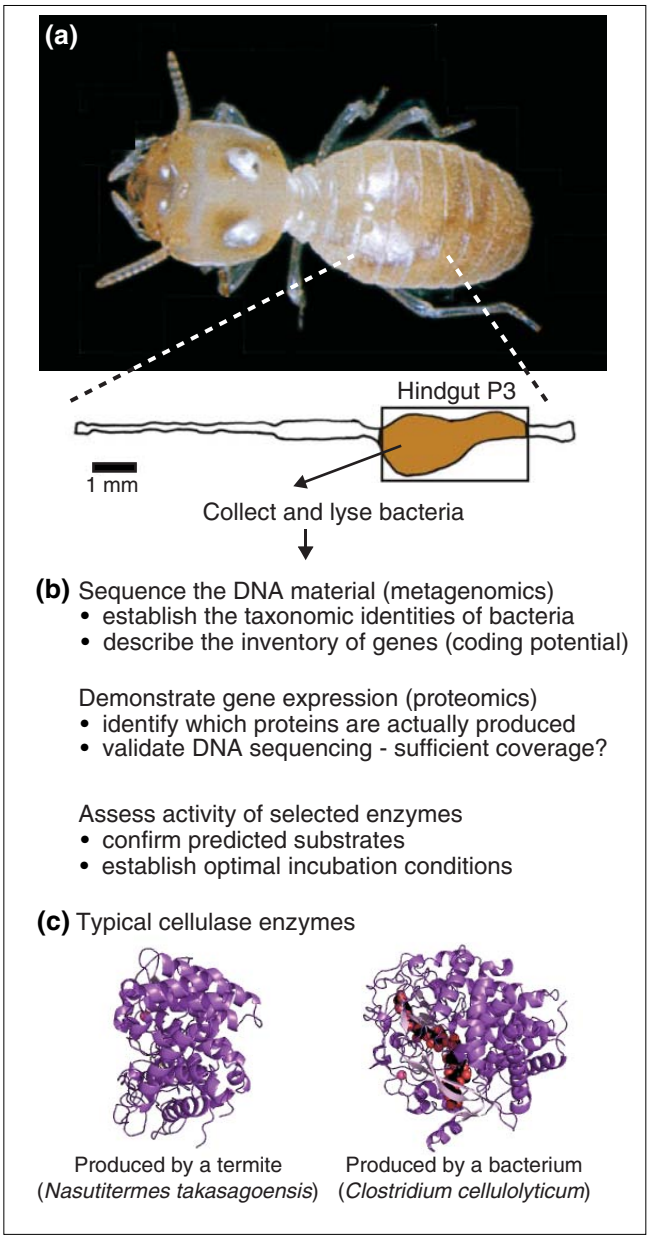


Figure 1
Exploring the termite hindgut. **(a)** Photograph of a worker termite from the genus Nasutitermes. **(b)** The gut contents from the third proctodeal segment (P3) were sampled, and analyzed using a variety of techniques. **(c)** Three-dimensional structures of two typical cellulase enzymes (left, PDB1ksd; right, PDB1f9d). Photograph: CSIRO.

biochemical analysis of some of the enzymes they discovered.

The team began by sampling the luminal contents of the P3 hindgut segment, pooling the material from 165 adult worker termites. This is the largest of the gut compartments, yet still contains only about 1 μl of material (Figure 1). From this material, the authors purified the genomic DNA, fragmented, subcloned and sequenced it. They generated about 70 megabases (Mb) of raw shotgun sequence and also selected several fosmid inserts to be sequenced separately for more detailed inspection. Warnecke *et al.* [3] mainly used classical capillary sequencing; today, this technology is being rapidly surpassed and the next-generation sequencing technologies will increase the scope of such studies by orders of magnitude [10]. As is the case for most metagenomics projects, the shotgun reads could not be assembled into complete genomes. In fact, relatively little assembly was possible at all - the longest assembled contig encompasses a mere 14.7 kb - owing to the complexity of the microbial community.

The metagenomics sequencing effort was complemented by a more focused strategy to sample a single phylogenetic marker gene (using PCR amplification and cloning of 16S ribosomal RNA genes). These 16S sequences were combined with similar sequences from the shotgun approach and analyzed in order to ask the question: which phyla and how many species are present in the termite gut?

As previously reported, members of the bacterial phyla Spirochaetes and Fibrobacteres dominated the community. Notably, Warnecke *et al.* [3] did not detect any archaeal sequences, nor did they find much eukaryotic material (there was apparently very little contaminating DNA from the host, if any). They discovered 216 distinct 'phylotypes' of bacteria (that is, groups of 16S sequences with at least 99% sequence identity) and estimated from the redundancy in these phylotypes that what they had found represented about 70-90% of the total diversity. This is roughly similar to the diversity of the human gut microbial flora [11].

Apart from a phylogenetic characterization, the authors carried out a quantitative analysis of functional genes in the sample. They focused on certain categories of interest: how many genes would encode enzymes known to degrade cellulose, xylan or lignin? Would there be evidence for nitrogen fixation? To find out, the authors grouped the predicted genes into families and orthologous groups, annotated them, and compared the abundance of each gene family to the respective occurrences of these genes in other environments, such as soil [7], seawater [6] or the human gut [12].

First and foremost, they found a large number of glycoside hydrolases; that is, enzymes that can degrade polysaccharides. The authors classified these genes according to known sequence families and predicted substrates, and attempted to assign them to the most likely source organism. Forty-five distinct groups were detected, and composition-based analysis predicted *Treponema* (a genus of Spirochaetes) as the most likely origin for the majority of these enzymes. In addition, a number of gene families known to associate with glycoside hydrolase

domains were found, including carbohydrate-binding domains and other functional domains. In total, hundreds of new enzymes were described, many of which significantly extend our knowledge of the various enzyme families. Remarkably, no enzymes were found for the degradation of lignin, a major constituent of wood that is partly responsible for its strength. Some enzymes capable of lignin degradation have previously been described, but none of these was found among the sequences retrieved here. Of course, as yet undescribed enzymes could do the task, or this activity could be located in a different compartment of the termite gut. The latter might well be the case, as many of the enzymes known to degrade lignin require molecular oxygen and the P3 segment is largely anoxic.

As expected, several other functional processes known (or suspected) to be carried out by the gut microbes were represented among the sequences. These include nitrogen fixation, chemotaxis and chemosensation, as well as carbon fixation from carbon dioxide via the Wood-Ljungdahl pathway [13].

Metaproteomics and activity assays

The detection of an open reading frame alone does not suffice to show that the protein is actually made, nor does it readily indicate when and where the gene is expressed. To assess the more abundant proteins at least, mass spectrometry is a promising tool, provided that the community is not too complex and it has been sampled deeply enough at the nucleotide level [9].

Warnecke and co-workers [3] have focused on a particular subset of the proteome (the secreted extracellular proteins) by analyzing centrifuged and clarified P3 luminal fluid using mass spectrometry. Although they were able to detect only a relatively small fraction of the expected proteins, they confirmed for the first time that bacterial glycosidases are indeed produced in the termite gut. What is more, they actually demonstrated activity for a number of these enzymes. More than 40 of the glycosidase genes were individually cloned, expressed heterologously and tested on acid-solubilized and microcrystalline cellulose. Although this is unlikely to match the situation in which these genes work in vivo, it shows convincingly that termite guts harbor secreted functional glycosidase enzymes.

Who encodes what?

The most pressing question in any metagenomics analysis is to what extent the molecular functions identified can be assigned to particular microbial lineages. This information is still almost entirely lacking for all but the simplest microbial communities, but it is crucially important for any deeper understanding of the ecology of these assemblages. The problem remains largely unsolved: in the current study [3], compositional analysis of the DNA provided classification for only 9% of the contigs beyond the phylum level, leading to uncertainties; for example, none of the *nifH* nitrogen-fixation genes could be assigned. Even where it does work, compositional analysis is probably not very reliable, as microbial genomes can harbor large stretches of recently acquired genetic material, which may not yet have equilibrated with the host genome. For individual genes of interest, clever use of coupled PCR reactions has recently shown a way to reliably map genes to their host genomes [14], but for a global solution we will probably have to wait for single-cell sequencing [15].

One of the most intriguing results of this study actually concerns a class of proteins to which no molecular function can be assigned so far. Warnecke *et al*. [3] identified a number of previously uncharacterized protein families that were strongly enriched compared with other metagenomes, and that were in some cases even quite specific to the termite gut microbes. This is exciting because the degradation of lignocellulose in most cases requires not just individual enzymes operating in isolation, but large macromolecular complexes that guide and coordinate the process. These complexes have been termed 'cellulosomes' and are (partially) known for a number of microbial species [16]. Scaffold proteins and accessory proteins may, however, be different from lineage to lineage, and this could mean that a number of unknown cellulosome-like proteins are contained in the specifically enriched proteins discovered in this study.

As an aside, we hope that the success of the gene-based approaches illustrated here and elsewhere will not deter those who seek to characterize individual microbial lineages more thoroughly. Isolating and growing microbes in pure culture remains an art, and one that continues to produce ground-breaking insights [17-19]. It provides unequivocal anchors for taxonomists and for functional studies, and allows access to the slow-growing, rare community members that can contribute essential functions. Comparative genomicists depend on a continued input of high-quality, well annotated genome sequences to sort out phylogeny and to understand the effects of horizontal gene transfers and other evolutionary phenomena. It is to be hoped that those who produce isolates and complete genome sequences will continue to be given appropriate credit for their work.

Wood as a source of biofuel

Can the results of this study help us make better use of wood as a fuel? Humans have used wood as an energy source for thousands of years, mostly for domestic heating and cooking. But it has also been used to generate power, for example in steam engines and occasionally by converting it to fuel for use in combustion engines (Figure 2). Conversion of wood into a biofuel, such as ethanol, is again a hot topic [20,21] because of its potential for at least partially replacing

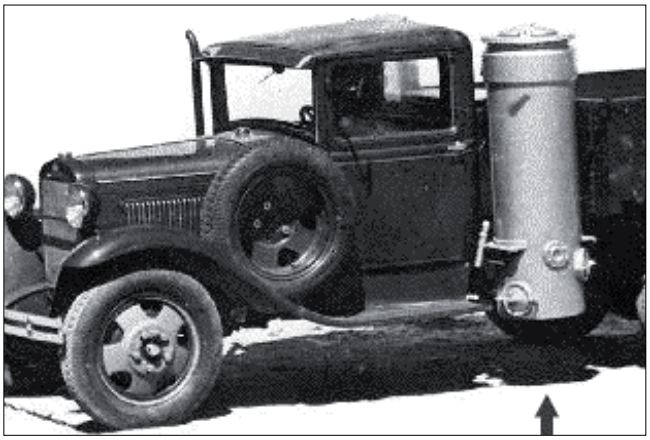


Figure 2
Making fuel from wood. The photograph, taken in 1951, shows a Russian automobile fitted with a 'wood gasifier' (arrow). Similar vehicles were relatively widespread in Europe in the 1940s and 50s, and achieved conversion efficiencies of roughly 3 kg of wood consumed per power-output equivalent to 1 liter of gasoline. Modern biotechnological approaches, using enzymes like the ones found in termite guts, are still struggling to surpass that efficiency [20]. But they do offer a much more convenient and clean fuel product, ethanol.

fossil fuels in transportation and thereby lowering greenhouse gas emissions.

Unlike some first-generation biofuels derived from just a small, energy-rich part of the plant (such as the seeds), which have been criticized on environmental grounds [22], wood-based biofuels use almost the whole plant. Trees in particular seem suitable for biofuel production, as they can be grown on marginal soils with very little water or fertilizer and do not compete with food crops.

Today, wood conversion is being attempted on the industrial scale using biotechnology. Cellulases and hemicellulases are already being used in this process and these enzymes can be further optimized. Many bigger challenges remain: how best to deal with the lignin, how best to pre-treat the wood and how to more efficiently release all sugars for fermentation. As termites achieve all of that in a volume of 1 μl, and at ambient temperatures, it seems that we have a lot to learn from them. It would be very satisfying if basic research into termite physiology could ultimately end up helping us to make better, environmentally friendly fuels.

Acknowledgements

The authors acknowledge support from the University of Zurich, through its research priority program 'Systems Biology and Functional Genomics'.

References

1. Breznak JA, Brune A: **Role of microorganisms in the digestion of lignocellulose by termites.** *Annu Rev Entomol* 1994, **39:**453-487.
2. Lilburn TG, Kim KS, Ostrom NE, Byzek KR, Leadbetter JR, Breznak JA: **Nitrogen fixation by symbiotic and free-living Spirochaetes.** *Science* 2001, **292:**2495-2498.
3. Warnecke F, Luginbühl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, *et al.*: **Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite.** *Nature* 2007, **450:**560-565.
4. Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, *et al.*: **Bacterial rhodopsin: evidence for a new type of phototrophy in the sea.** *Science* 2000, **289:**1902-1906.
5. Treusch AH, Leininger S, Kletzin A, Schuster SC, Klenk HP, Schleper C: **Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling.** *Environ Microbiol* 2005, **7:**1985-1995.
6. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, *et al.*: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304:**66-74.
7. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, *et al.*: **Comparative metagenomics of microbial communities.** *Science* 2005, **308:**554-557.
8. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment.** *Nature* 2004, **428:**37-43.
9. Ram RJ, Verberkmoes NC, Thelen MP, Tyson GW, Baker BJ, Blake RC, 2nd, Shah M, Hettich RL, Banfield JF: **Community proteomics of a natural microbial biofilm.** *Science* 2005, **308:**1915-1920.
10. Ryan D, Rahimi M, Lund J, Mehta R, Parviz BA: **Toward nanoscale genome sequencing.** *Trends Biotechnol* 2007, **25:**385-389.
11. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA: **Diversity of the human intestinal microbial flora.** *Science* 2005, **308:**1635-1638.
12. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome.** *Science* 2006, **312:**1355-1359.
13. Graber JR, Breznak JA: **Physiology and nutrition of *Treponema primitia*, an H2/CO2-acetogenic Spirochaete from termite hindguts.** *Appl Environ Microbiol* 2004, **70:**1307-1314.
14. Ottesen EA, Hong JW, Quake SR, Leadbetter JR: **Microfluidic digital PCR enables multigene analysis of individual environmental bacteria.** *Science* 2006, **314:**1464-1467.
15. Lasken RS: **Single-cell genomic sequencing using multiple displacement amplification.** *Curr Opin Microbiol* 2007, **10:**510-516.
16. Doi RH, Kosugi A: **Cellulosomes: plant-cell-wall-degrading enzyme complexes.** *Nat Rev Microbiol* 2004, **2:**541-551.
17. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, *et al.*: **Genome streamlining in a cosmopolitan oceanic bacterium.** *Science* 2005, **309:**1242-1245.
18. Giraud E, Moulin L, Vallenet D, Barbe V, Cytryn E, Avarre JC, Jaubert M, Simon D, Cartieaux F, Prin Y, *et al.*: **Legumes symbioses: absence of Nod genes in photosynthetic bradyrhizobia.** *Science* 2007, **316:**1307-1312.
19. Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, Barnstead M, Beeson KY, Bibbs L, Bolanos R, Keller M, *et al.*: **The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism.** *Proc Natl Acad Sci USA* 2003, **100:**12984-12988.
20. Wyman CE: **What is (and is not) vital to advancing cellulosic ethanol.** *Trends Biotechnol* 2007, **25:**153-157.
21. Hahn-Hagerdal B, Galbe M, Gorwa-Grauslund MF, Liden G, Zacchi G: **Bio-ethanol - the fuel of tomorrow from the residues of today.** *Trends Biotechnol* 2006, **24:**549-556.
22. Pimentel D, Patzek T, Cecil G: **Ethanol production: energy, economic, and environmental losses.** *Rev Environ Contam Toxicol* 2007, **189:**25-41.

BIBLIOGRAPHY

[1] Mark Achtman and Michael Wagner. Microbial diversity and the genetic nature of microbial species. *Nature Reviews. Microbiology*, 6(6):431–440, June 2008. ISSN 1740-1534. doi: 10.1038/nrmicro1872. URL http://www.ncbi.nlm.nih.gov/pubmed/18461076. PMID: 18461076.

[2] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, March 2003. ISSN 0028-0836. doi: 10.1038/nature01511. URL http://www.ncbi.nlm.nih.gov/pubmed/12634793. PMID: 12634793.

[3] R I Amann, L Krumholz, and D A Stahl. Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. *Journal of Bacteriology*, 172(2):762–770, February 1990. ISSN 0021-9193. URL http://www.ncbi.nlm.nih.gov/pubmed/1688842. PMID: 1688842.

[4] Fredrik Bäckhed, Ruth E Ley, Justin L Sonnenburg, Daniel A Peterson, and Jeffrey I Gordon. Host-bacterial mutualism in the human intestine. *Science (New York, N.Y.)*, 307(5717):1915–1920, March 2005. ISSN 1095-9203. doi: 10.1126/science.1104816. URL http://www.ncbi.nlm.nih.gov/pubmed/15790844. PMID: 15790844.

[5] Brett J Baker, Gene W Tyson, Richard I Webb, Judith Flanagan, Philip Hugenholtz, Eric E Allen, and Jillian F Banfield. Lineages of acidophilic archaea revealed by community genomic analysis. *Science (New York, N.Y.)*, 314(5807):1933–1935, December 2006. ISSN 1095-9203. doi: 10.1126/science.1132690. URL http://www.ncbi.nlm.nih.gov/pubmed/17185602. PMID: 17185602.

[6] G C Baker, J J Smith, and D A Cowan. Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods*, 55(3):541–555, December 2003. ISSN 0167-7012. URL http://www.ncbi.nlm.nih.gov/pubmed/14607398. PMID: 14607398.

[7] Richard D Bardgett, Chris Freeman, and Nicholas J Ostle. Microbial contributions to climate change through carbon cycle feedbacks. *The ISME Journal*, 2(8):805–814, August 2008. ISSN 1751-7370. doi: 10.1038/ismej.2008.58. URL http://www.ncbi.nlm.nih.gov/pubmed/18615117. PMID: 18615117.

[8] O Béjá, L Aravind, E V Koonin, M T Suzuki, A Hadd, L P Nguyen, S B Jovanovich, C M Gates, R A Feldman, J L Spudich, E N Spudich, and E F DeLong. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science (New York, N.Y.)*, 289(5486):1902–1906, September 2000. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/10988064. PMID: 10988064.

[9] O Béjá, M T Suzuki, E V Koonin, L Aravind, A Hadd, L P Nguyen, R Villacorta, M Amjadi, C Garrigues, S B Jovanovich, R A Feldman, and E F DeLong.

Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environmental Microbiology*, 2(5):516–529, October 2000. ISSN 1462-2912. URL http://www.ncbi.nlm.nih.gov/pubmed/11233160. PMID: 11233160.

[10] O Béjá, E N Spudich, J L Spudich, M Leclerc, and E F DeLong. Proteorhodopsin phototrophy in the ocean. *Nature*, 411(6839):786–789, June 2001. ISSN 0028-0836. doi: 10.1038/35081051. URL http://www.ncbi.nlm.nih.gov/pubmed/11459054. PMID: 11459054.

[11] Ido Braslavsky, Benedict Hebert, Emil Kartalov, and Stephen R Quake. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7):3960–3964, April 2003. ISSN 0027-8424. doi: 10.1073/pnas.0230489100. URL http://www.ncbi.nlm.nih.gov/pubmed/12651960. PMID: 12651960.

[12] T D Brock and M L Brock. Autoradiography as a tool in microbial ecology. *Nature*, 209(5024):734–736, February 1966. ISSN 0028-0836. URL http://www.ncbi.nlm.nih.gov/pubmed/5922142. PMID: 5922142.

[13] Michael A. Brockhurst, Andrew D. Morgan, Paul B. Rainey, and Angus Buckling. Population mixing accelerates coevolution. *Ecology Letters*, 6(11):975–979, 2003. doi: 10.1046/j.1461-0248.2003.00531.x. URL http://dx.doi.org/10.1046/j.1461-0248.2003.00531.x.

[14] J Brosius, M L Palmer, P J Kennedy, and H F Noller. Complete nucleotide sequence of a 16S ribosomal RNA gene from escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 75(10):4801–4805, October 1978. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/368799. PMID: 368799.

[15] Erich Brunner, Christian H Ahrens, Sonali Mohanty, Hansruedi Baetschmann, Sandra Loevenich, Frank Potthast, Eric W Deutsch, Christian Panse, Ulrik de Lichtenberg, Oliver Rinner, Hookeun Lee, Patrick G A Pedrioli, Johan Malmstrom, Katja Koehler, Sabine Schrimpf, Jeroen Krijgsveld, Floyd Kregenow, Albert J R Heck, Ernst Hafen, Ralph Schlapbach, and Ruedi Aebersold. A high-quality catalog of the drosophila melanogaster proteome. *Nature Biotechnology*, 25(5):576–583, May 2007. ISSN 1087-0156. doi: 10.1038/nbt1300. URL http://www.ncbi.nlm.nih.gov/pubmed/17450130. PMID: 17450130.

[16] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pe na, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, May 2010. ISSN 1548-7105. doi: 10.1038/nmeth.f.303. URL http://www.ncbi.nlm.nih.gov/pubmed/20383131. PMID: 20383131.

[17] Samuel Chaffron and Christian von Mering. Termites in the woodwork. *Genome Biology*, 8(11):229, 2007. ISSN 1465-6914. doi: 10.1186/gb-2007-8-11-229. URL http://www.ncbi.nlm.nih.gov/pubmed/18036268. PMID: 18036268.

[18] Samuel Chaffron, Hubert Rehrauer, Jakob Pernthaler, and Christian von Mering. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, June 2010. ISSN 1549-5469. doi: 10.1101/gr.104521.109. URL http://www.ncbi.nlm.nih.gov/pubmed/20458099. PMID: 20458099.

[19] J R Cole, Q Wang, E Cardenas, J Fish, B Chai, R J Farris, A S Kulam-Syed-Mohideen, D M McGarrell, T Marsh, G M Garrity, and J M Tiedje. The ribosomal database project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37(Database issue):D141–145, January 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn879. URL http://www.ncbi.nlm.nih.gov/pubmed/19004872. PMID: 19004872.

[20] Frank B Dean, Seiyu Hosono, Linhua Fang, Xiaohong Wu, A Fawad Faruqi, Patricia Bray-Ward, Zhenyu Sun, Qiuling Zong, Yuefen Du, Jing Du, Mark Driscoll, Wanmin Song, Stephen F Kingsmore, Michael Egholm, and Roger S Lasken. Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8):5261–5266, April 2002. ISSN 0027-8424. doi: 10.1073/pnas.082089499. URL http://www.ncbi.nlm.nih.gov/pubmed/11959976. PMID: 11959976.

[21] Anne E Dekas, Rachel S Poretsky, and Victoria J Orphan. Deep-sea archaea fix and share nitrogen in methane-consuming microbial consortia. *Science (New York, N.Y.)*, 326(5951):422–426, October 2009. ISSN 1095-9203. doi: 10.1126/science.1178223. URL http://www.ncbi.nlm.nih.gov/pubmed/19833965. PMID: 19833965.

[22] Nathanaël Delmotte, Claudia Knief, Samuel Chaffron, Gerd Innerebner, Bernd Roschitzki, Ralph Schlapbach, Christian von Mering, and Julia A Vorholt. Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 106(38):16428–16433, September 2009. ISSN 1091-6490. doi: 10.1073/pnas.0905240106. URL http://www.ncbi.nlm.nih.gov/pubmed/19805315. PMID: 19805315.

[23] E F DeLong. Microbial seascapes revisited. *Current Opinion in Microbiology*, 4(3):290–295, June 2001. ISSN 1369-5274. URL http://www.ncbi.nlm.nih.gov/pubmed/11378481. PMID: 11378481.

[24] E F DeLong, G S Wickham, and N R Pace. Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. *Science (New York, N.Y.)*, 243(4896):1360–1363, March 1989. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/2466341. PMID: 2466341.

[25] Vincent J Denef, Nathan C VerBerkmoes, Manesh B Shah, Paul Abraham, Mark Lefsrud, Robert L Hettich, and Jillian F Banfield. Proteomics-inferred genome typing (PIGT) demonstrates inter-population recombination as a strategy for environmental adaptation. *Environmental Microbiology*, 11(2):313–325, February 2009. ISSN 1462-2920. doi: 10.1111/j.1462-2920.2008.01769.x. URL http://www.ncbi.nlm.nih.gov/pubmed/18826438. PMID: 18826438.

[26] Vincent J Denef, Linda H Kalnejais, Ryan S Mueller, Paul Wilmes, Brett J Baker, Brian C Thomas, Nathan C VerBerkmoes, Robert L Hettich, and Jillian F Banfield. Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proceedings of the National Academy of Sciences of the United States of America*, 107(6):2383–2390, February 2010. ISSN 1091-6490. doi: 10.1073/pnas.0907041107. URL http://www.ncbi.nlm.nih.gov/pubmed/20133593. PMID: 20133593.

[27] T Z DeSantis, P Hugenholtz, N Larsen, M Rojas, E L Brodie, K Keller, T Huber, D Dalevi, P Hu, and G L Andersen. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, 72(7):5069–5072, July 2006. ISSN 0099-2240. doi: 10.1128/AEM.03006-05. URL http://www.ncbi.nlm.nih.gov/pubmed/16820507. PMID: 16820507.

[28] Frank Desiere, Eric W Deutsch, Alexey I Nesvizhskii, Parag Mallick, Nichole L King, Jimmy K Eng, Alan Aderem, Rose Boyle, Erich Brunner, Samuel Donohoe, Nelson Fausto, Ernst Hafen, Lee Hood, Michael G Katze, Kathleen A Kennedy, Floyd Kregenow, Hookeun Lee, Biaoyang Lin, Dan Martin, Jeffrey A Ranish, David J Rawlings, Lawrence E Samelson, Yuzuru Shiio, Julian D Watts, Bernd Wollscheid, Michael E Wright, Wei Yan, Lihong Yang, Eugene C Yi, Hui Zhang, and Ruedi Aebersold. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biology*, 6(1):R9, 2005. ISSN 1465-6914. doi: 10.1186/gb-2004-6-1-r9. URL http://www.ncbi.nlm.nih.gov/pubmed/15642101. PMID: 15642101.

[29] Les Dethlefsen, Margaret McFall-Ngai, and David A Relman. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature*, 449(7164):811–818, October 2007. ISSN 1476-4687. doi: 10.1038/nature06245. URL http://www.ncbi.nlm.nih.gov/pubmed/17943117. PMID: 17943117.

[30] Les Dethlefsen, Sue Huse, Mitchell L Sogin, and David A Relman. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biology*, 6(11):e280, November 2008. ISSN 1545-7885. doi: 10.1371/journal.pbio.0060280. URL http://www.ncbi.nlm.nih.gov/pubmed/19018661. PMID: 19018661.

[31] Elizabeth A Dinsdale, Robert A Edwards, Dana Hall, Florent Angly, Mya Breitbart, Jennifer M Brulc, Mike Furlan, Christelle Desnues, Matthew Haynes, Linlin Li, Lauren McDaniel, Mary Ann Moran, Karen E Nelson, Christina Nilsson, Robert Olson, John Paul, Beltran Rodriguez Brito, Yijun Ruan, Brandon K Swan, Rick Stevens, David L Valentine, Rebecca Vega Thurber, Linda

Wegley, Bryan A White, and Forest Rohwer. Functional metagenomic profiling of nine biomes. *Nature*, 452(7187):629–632, April 2008. ISSN 1476-4687. doi: 10.1038/nature06810. URL http://www.ncbi.nlm.nih.gov/pubmed/18337718. PMID: 18337718.

[32] W Ford Doolittle and Olga Zhaxybayeva. On the origin of prokaryotic species. *Genome Research*, 19(5):744–756, May 2009. ISSN 1088-9051. doi: 10.1101/gr.086645.108. URL http://www.ncbi.nlm.nih.gov/pubmed/19411599. PMID: 19411599.

[33] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex Dewinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, 323(5910):133–138, January 2009. ISSN 1095-9203. doi: 10.1126/science.1162986. URL http://www.ncbi.nlm.nih.gov/pubmed/19023044. PMID: 19023044.

[34] Hossein Fakhrai-Rad, Nader Pourmand, and Mostafa Ronaghi. Pyrosequencing: an accurate detection platform for single nucleotide polymorphisms. *Human Mutation*, 19(5):479–485, May 2002. ISSN 1098-1004. doi: 10.1002/humu.10078. URL http://www.ncbi.nlm.nih.gov/pubmed/11968080. PMID: 11968080.

[35] Paul G Falkowski, Tom Fenchel, and Edward F Delong. The microbial engines that drive earth's biogeochemical cycles. *Science (New York, N.Y.)*, 320(5879):1034–1039, May 2008. ISSN 1095-9203. doi: 10.1126/science.1153213. URL http://www.ncbi.nlm.nih.gov/pubmed/18497287. PMID: 18497287.

[36] Edward J Feil. Small change: keeping pace with microevolution. *Nature Reviews. Microbiology*, 2(6):483–495, June 2004. ISSN 1740-1526. doi: 10.1038/nrmicro904. URL http://www.ncbi.nlm.nih.gov/pubmed/15152204. PMID: 15152204.

[37] R D Fleischmann, M D Adams, O White, R A Clayton, E F Kirkness, A R Kerlavage, C J Bult, J F Tomb, B A Dougherty, and J M Merrick. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science (New York, N.Y.)*, 269(5223):496–512, July 1995. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/7542800. PMID: 7542800.

[38] C M Fraser, J D Gocayne, O White, M D Adams, R A Clayton, R D Fleischmann, C J Bult, A R Kerlavage, G Sutton, J M Kelley, R D Fritchman, J F Weidman, K V Small, M Sandusky, J Fuhrmann, D Nguyen, T R Utterback, D M Saudek, C A Phillips, J M Merrick, J F Tomb, B A Dougherty, K F Bott, P C Hu, T S Lucier, S N Peterson, H O Smith, C A Hutchison, and J C Venter. The minimal gene

complement of mycoplasma genitalium. *Science (New York, N.Y.)*, 270(5235): 397–403, October 1995. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/7569993. PMID: 7569993.

[39] Jorge Frias-Lopez, Yanmei Shi, Gene W Tyson, Maureen L Coleman, Stephan C Schuster, Sallie W Chisholm, and Edward F Delong. Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences of the United States of America*, 105(10):3805–3810, March 2008. ISSN 1091-6490. doi: 10.1073/pnas.0708897105. URL http://www.ncbi.nlm.nih.gov/pubmed/18316740. PMID: 18316740.

[40] Laura S Frost, Raphael Leplae, Anne O Summers, and Ariane Toussaint. Mobile genetic elements: the agents of open source evolution. *Nature Reviews. Microbiology*, 3(9):722–732, September 2005. ISSN 1740-1526. doi: 10.1038/nrmicro1235. URL http://www.ncbi.nlm.nih.gov/pubmed/16138100. PMID: 16138100.

[41] J A Fuhrman, K McCallum, and A A Davis. Novel major archaebacterial group from marine plankton. *Nature*, 356(6365):148–149, March 1992. ISSN 0028-0836. doi: 10.1038/356148a0. URL http://www.ncbi.nlm.nih.gov/pubmed/1545865. PMID: 1545865.

[42] Jed A Fuhrman. Microbial community structure and its functional implications. *Nature*, 459(7244):193–199, May 2009. ISSN 1476-4687. doi: 10.1038/nature08058. URL http://www.ncbi.nlm.nih.gov/pubmed/19444205. PMID: 19444205.

[43] Jed A Fuhrman, Ian Hewson, Michael S Schwalbach, Joshua A Steele, Mark V Brown, and Shahid Naeem. Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences of the United States of America*, 103(35):13104–13109, August 2006. ISSN 0027-8424. doi: 10.1073/pnas.0602399103. URL http://www.ncbi.nlm.nih.gov/pubmed/16938845. PMID: 16938845.

[44] Dirk Gevers, Frederick M Cohan, Jeffrey G Lawrence, Brian G Spratt, Tom Coenye, Edward J Feil, Erko Stackebrandt, Yves Van de Peer, Peter Vandamme, Fabiano L Thompson, and Jean Swings. Opinion: Re-evaluating prokaryotic species. *Nature Reviews. Microbiology*, 3(9):733–739, September 2005. ISSN 1740-1526. doi: 10.1038/nrmicro1236. URL http://www.ncbi.nlm.nih.gov/pubmed/16138101. PMID: 16138101.

[45] S J Giovannoni, T B Britschgi, C L Moyer, and K G Field. Genetic diversity in sargasso sea bacterioplankton. *Nature*, 345(6270):60–63, May 1990. ISSN 0028-0836. doi: 10.1038/345060a0. URL http://www.ncbi.nlm.nih.gov/pubmed/2330053. PMID: 2330053.

[46] William J Greenleaf and Steven M Block. Single-molecule, motion-based DNA sequencing using RNA polymerase. *Science (New York, N.Y.)*, 313(5788): 801, August 2006. ISSN 1095-9203. doi: 10.1126/science.1130105. URL http://www.ncbi.nlm.nih.gov/pubmed/16902131. PMID: 16902131.

[47] Kevin Gross. Positive interactions among competitors can produce species-rich communities. *Ecology Letters*, 11(9):929–936, September 2008. ISSN 1461-0248.

doi: 10.1111/j.1461-0248.2008.01204.x. URL http://www.ncbi.nlm.nih.gov/pubmed/18485001. PMID: 18485001.

[48] D S Guttman and D E Dykhuizen. Clonal divergence in escherichia coli as a result of recombination, not mutation. *Science (New York, N.Y.)*, 266(5189): 1380–1383, November 1994. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/7973728. PMID: 7973728.

[49] Steven J Hallam, Nik Putnam, Christina M Preston, John C Detter, Daniel Rokhsar, Paul M Richardson, and Edward F DeLong. Reverse methanogenesis: testing the hypothesis with environmental genomics. *Science (New York, N.Y.)*, 305(5689):1457–1462, September 2004. ISSN 1095-9203. doi: 10.1126/science.1100025. URL http://www.ncbi.nlm.nih.gov/pubmed/15353801. PMID: 15353801.

[50] Micah Hamady, Catherine Lozupone, and Rob Knight. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *The ISME Journal*, 4(1):17–27, January 2010. ISSN 1751-7370. doi: 10.1038/ismej.2009.97. URL http://www.ncbi.nlm.nih.gov/pubmed/19710709. PMID: 19710709.

[51] Timothy D Harris, Phillip R Buzby, Hazen Babcock, Eric Beer, Jayson Bowers, Ido Braslavsky, Marie Causey, Jennifer Colonell, James Dimeo, J William Efcavitch, Eldar Giladi, Jaime Gill, John Healy, Mirna Jarosz, Dan Lapen, Keith Moulton, Stephen R Quake, Kathleen Steinmann, Edward Thayer, Anastasia Tyurina, Rebecca Ward, Howard Weiss, and Zheng Xie. Single-molecule DNA sequencing of a viral genome. *Science (New York, N.Y.)*, 320(5872): 106–109, April 2008. ISSN 1095-9203. doi: 10.1126/science.1150427. URL http://www.ncbi.nlm.nih.gov/pubmed/18388294. PMID: 18388294.

[52] R David Hawkins, Gary C Hon, and Bing Ren. Next-generation genomics: an integrative approach. *Nature Reviews. Genetics*, June 2010. ISSN 1471-0064. doi: 10.1038/nrg2795. URL http://www.ncbi.nlm.nih.gov/pubmed/20531367. PMID: 20531367.

[53] Lynette Hirschman, Cheryl Clark, K Bretonnel Cohen, Scott Mardis, Joanne Luciano, Renzo Kottmann, James Cole, Victor Markowitz, Nikos Kyrpides, Norman Morrison, Lynn M Schriml, and Dawn Field. Habitat-Lite: a GSC case study based on free text terms for environmental metadata. *Omics: A Journal of Integrative Biology*, 12(2):129–136, June 2008. ISSN 1536-2310. doi: 10.1089/omi.2008.0016. URL http://www.ncbi.nlm.nih.gov/pubmed/18416669. PMID: 18416669.

[54] Philip Hugenholtz. Exploring prokaryotic diversity in the genomic era. *Genome Biology*, 3(2):REVIEWS0003, 2002. ISSN 1465-6914. URL http://www.ncbi.nlm.nih.gov/pubmed/11864374. PMID: 11864374.

[55] J Huisman and F J Weissing. Fundamental unpredictability in multispecies competition. *The American Naturalist*, 157(5):488–494, May 2001. ISSN 1537-5323. doi: 10.1086/319929. URL http://www.ncbi.nlm.nih.gov/pubmed/18707257. PMID: 18707257.

[56] Souichiro Kato, Shin Haruta, Zong Jun Cui, Masaharu Ishii, and Yasuo Igarashi. Stable coexistence of five bacterial strains as a cellulose-degrading community. *Applied and Environmental Microbiology*, 71(11):7099–7106, November 2005. ISSN 0099-2240. doi: 10.1128/AEM.71.11.7099-7106.2005. URL http://www.ncbi.nlm.nih.gov/pubmed/16269746. PMID: 16269746.

[57] Martin Krzywinski, Jacqueline Schein, InanÃ§ Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, September 2009. ISSN 1549-5469. doi: 10.1101/gr.092759.109. URL http://www.ncbi.nlm.nih.gov/pubmed/19541911. PMID: 19541911.

[58] Victor Kunin, Anna Engelbrektson, Howard Ochman, and Philip Hugenholtz. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, 12(1):118–123, January 2010. ISSN 1462-2920. doi: 10.1111/j.1462-2920.2009.02051.x. URL http://www.ncbi.nlm.nih.gov/pubmed/19725865. PMID: 19725865.

[59] J G Lawrence. Gene transfer, speciation, and the evolution of bacterial genomes. *Current Opinion in Microbiology*, 2(5):519–523, October 1999. ISSN 1369-5274. URL http://www.ncbi.nlm.nih.gov/pubmed/10508729. PMID: 10508729.

[60] Jeffrey G Lawrence. Gene transfer in bacteria: speciation without species? *Theoretical Population Biology*, 61(4):449–460, June 2002. ISSN 0040-5809. URL http://www.ncbi.nlm.nih.gov/pubmed/12167364. PMID: 12167364.

[61] Ruth E Ley, Micah Hamady, Catherine Lozupone, Peter J Turnbaugh, Rob Roy Ramey, J Stephen Bircher, Michael L Schlegel, Tammy A Tucker, Mark D Schrenzel, Rob Knight, and Jeffrey I Gordon. Evolution of mammals and their gut microbes. *Science (New York, N.Y.)*, 320(5883):1647–1651, June 2008. ISSN 1095-9203. doi: 10.1126/science.1155725. URL http://www.ncbi.nlm.nih.gov/pubmed/18497261. PMID: 18497261.

[62] Ian Lo, Vincent J Denef, Nathan C Verberkmoes, Manesh B Shah, Daniela Goltsman, Genevieve DiBartolo, Gene W Tyson, Eric E Allen, Rachna J Ram, J Chris Detter, Paul Richardson, Michael P Thelen, Robert L Hettich, and Jillian F Banfield. Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature*, 446(7135):537–541, March 2007. ISSN 1476-4687. doi: 10.1038/nature05624. URL http://www.ncbi.nlm.nih.gov/pubmed/17344860. PMID: 17344860.

[63] J M Logsdon and D M Faguy. Thermotoga heats up lateral gene transfer. *Current Biology: CB*, 9(19):R747–751, October 1999. ISSN 0960-9822. URL http://www.ncbi.nlm.nih.gov/pubmed/10531001. PMID: 10531001.

[64] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Godwin, Wen He, Scott Helgesen, Chun Heen Ho, Chun He Ho, Gerard P Irzyk, Szilveszter C Jando, Maria L I Alenquer, Thomas P Jarvie, Kshama B Jirage, Jong-Bum Kim,

James R Knight, Janna R Lanza, John H Leamon, Steven M Lefkowitz, Ming Lei, Jing Li, Kenton L Lohman, Hong Lu, Vinod B Makhijani, Keith E McDade, Michael P McKenna, Eugene W Myers, Elizabeth Nickerson, John R Nobile, Ramona Plant, Bernard P Puc, Michael T Ronan, George T Roth, Gary J Sarkis, Jan Fredrik Simons, John W Simpson, Maithreyan Srinivasan, Karrie R Tartaro, Alexander Tomasz, Kari A Vogt, Greg A Volkmer, Shally H Wang, Yong Wang, Michael P Weiner, Pengguang Yu, Richard F Begley, and Jonathan M Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, September 2005. ISSN 1476-4687. doi: 10.1038/nature03959. URL http://www.ncbi.nlm.nih.gov/pubmed/16056220. PMID: 16056220.

[65] E A Mayer and J P Baldi. Can regulatory peptides be regarded as words of a biological language. *The American Journal of Physiology*, 261(2 Pt 1):G171–184, August 1991. ISSN 0002-9513. URL http://www.ncbi.nlm.nih.gov/pubmed/1872391. PMID: 1872391.

[66] Margaret McFall-Ngai. Adaptive immunity: care for the community. *Nature*, 445(7124):153, January 2007. ISSN 1476-4687. doi: 10.1038/445153a. URL http://www.ncbi.nlm.nih.gov/pubmed/17215830. PMID: 17215830.

[67] Alice Carolyn McHardy, HÃ©ctor GarcÃa MartÃn, Aristotelis Tsirigos, Philip Hugenholtz, and Isidore Rigoutsos. Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, 4(1):63–72, January 2007. ISSN 1548-7091. doi: 10.1038/nmeth976. URL http://www.ncbi.nlm.nih.gov/pubmed/17179938. PMID: 17179938.

[68] Ted H M Mes and Marije Doeleman. Positive selection on transposase genes of insertion sequences in the crocosphaera watsonii genome. *Journal of Bacteriology*, 188(20):7176–7185, October 2006. ISSN 0021-9193. doi: 10.1128/JB.01021-06. URL http://www.ncbi.nlm.nih.gov/pubmed/17015656. PMID: 17015656.

[69] Ryan S Mueller, Vincent J Denef, Linda H Kalnejais, K Blake Suttle, Brian C Thomas, Paul Wilmes, Richard L Smith, D Kirk Nordstrom, R Blaine McCleskey, Manesh B Shah, Nathan C Verberkmoes, Robert L Hettich, and Jillian F Banfield. Ecological distribution and population physiology defined by proteomics in a natural microbial community. *Molecular Systems Biology*, 6:374, June 2010. ISSN 1744-4292. doi: 10.1038/msb.2010.30. URL http://www.ncbi.nlm.nih.gov/pubmed/20531404. PMID: 20531404.

[70] G Muyzer, E C de Waal, and A G Uitterlinden. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology*, 59(3):695–700, March 1993. ISSN 0099-2240. URL http://www.ncbi.nlm.nih.gov/pubmed/7683183. PMID: 7683183.

[71] K E Nelson, R A Clayton, S R Gill, M L Gwinn, R J Dodson, D H Haft, E K Hickey, J D Peterson, W C Nelson, K A Ketchum, L McDonald, T R Utterback, J A Malek, K D Linher, M M Garrett, A M Stewart, M D Cotton, M S Pratt, C A Phillips, D Richardson, J Heidelberg, G G Sutton, R D Fleischmann, J A Eisen, O White, S L Salzberg, H O Smith, J C Venter, and C M Fraser.

Evidence for lateral gene transfer between archaea and bacteria from genome sequence of thermotoga maritima. *Nature*, 399(6734):323–329, May 1999. ISSN 0028-0836. doi: 10.1038/20601. URL http://www.ncbi.nlm.nih.gov/pubmed/10360571. PMID: 10360571.

[72] Karen E Nelson, George M Weinstock, Sarah K Highlander, Kim C Worley, Heather Huot Creasy, Jennifer Russo Wortman, Douglas B Rusch, Makedonka Mitreva, Erica Sodergren, Asif T Chinwalla, Michael Feldgarden, Dirk Gevers, Brian J Haas, Ramana Madupu, Doyle V Ward, Bruce W Birren, Richard A Gibbs, Barbara Methe, Joseph F Petrosino, Robert L Strausberg, Granger G Sutton, Owen R White, Richard K Wilson, Scott Durkin, Michelle Gwinn Giglio, Sharvari Gujja, Clint Howarth, Chinnappa D Kodira, Nikos Kyrpides, Teena Mehta, Donna M Muzny, Matthew Pearson, Kymberlie Pepin, Amrita Pati, Xiang Qin, Chandri Yandava, Qiandong Zeng, Lan Zhang, Aaron M Berlin, Lei Chen, Theresa A Hepburn, Justin Johnson, Jamison McCorrison, Jason Miller, Pat Minx, Chad Nusbaum, Carsten Russ, Sean M Sykes, Chad M Tomlinson, Sarah Young, Wesley C Warren, Jonathan Badger, Jonathan Crabtree, Victor M Markowitz, Joshua Orvis, Andrew Cree, Steve Ferriera, Lucinda L Fulton, Robert S Fulton, Marcus Gillis, Lisa D Hemphill, Vandita Joshi, Christie Kovar, Manolito Torralba, Kris A Wetterstrand, Amr Abouellleil, Aye M Wollam, Christian J Buhay, Yan Ding, Shannon Dugan, Michael G FitzGerald, Mike Holder, Jessica Hostetler, Sandra W Clifton, Emma Allen-Vercoe, Ashlee M Earl, Candace N Farmer, Konstantinos Liolios, Michael G Surette, Qiang Xu, Craig Pohl, Katarzyna Wilczek-Boney, and Dianhui Zhu. A catalog of reference genomes from the human microbiome. *Science (New York, N.Y.)*, 328(5981): 994–999, May 2010. ISSN 1095-9203. doi: 10.1126/science.1183605. URL http://www.ncbi.nlm.nih.gov/pubmed/20489017. PMID: 20489017.

[73] Teresa Nogueira, Daniel J Rankin, Marie Touchon, FranÃ§ois Taddei, Sam P Brown, and Eduardo P C Rocha. Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Current Biology: CB*, 19(20):1683–1691, November 2009. ISSN 1879-0445. doi: 10.1016/j.cub.2009.08.056. URL http://www.ncbi.nlm.nih.gov/pubmed/19800234. PMID: 19800234.

[74] H Ochman, J G Lawrence, and E A Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, May 2000. ISSN 0028-0836. doi: 10.1038/35012500. URL http://www.ncbi.nlm.nih.gov/pubmed/10830951. PMID: 10830951.

[75] Howard Ochman and Liliana M Davalos. The nature and dynamics of bacterial genomes. *Science (New York, N.Y.)*, 311(5768):1730–1733, March 2006. ISSN 1095-9203. doi: 10.1126/science.1119966. URL http://www.ncbi.nlm.nih.gov/pubmed/16556833. PMID: 16556833.

[76] Victoria J Orphan. Methods for unveiling cryptic microbial partnerships in nature. *Current Opinion in Microbiology*, 12(3):231–237, June 2009. ISSN 1879-0364. doi: 10.1016/j.mib.2009.04.003. URL http://www.ncbi.nlm.nih.gov/pubmed/19447672. PMID: 19447672.

[77] Elizabeth A Ottesen, Jong Wook Hong, Stephen R Quake, and Jared R Leadbetter. Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. *Science (New York, N.Y.)*, 314(5804):1464–1467, December 2006. ISSN 1095-9203. doi: 10.1126/science.1131370. URL http://www.ncbi.nlm.nih.gov/pubmed/17138901. PMID: 17138901.

[78] Jörg Overmann and Karin Schubert. Phototrophic consortia: model systems for symbiotic interrelations between prokaryotes. *Archives of Microbiology*, 177(3): 201–208, March 2002. ISSN 0302-8933. doi: 10.1007/s00203-001-0377-z. URL http://www.ncbi.nlm.nih.gov/pubmed/11907675. PMID: 11907675.

[79] N R Pace. A molecular view of microbial diversity and the biosphere. *Science (New York, N.Y.)*, 276(5313):734–740, May 1997. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/9115194. PMID: 9115194.

[80] Chana Palmer, Elisabeth M Bik, Daniel B DiGiulio, David A Relman, and Patrick O Brown. Development of the human infant intestinal microbiota. *PLoS Biology*, 5(7):e177, July 2007. ISSN 1545-7885. doi: 10.1371/journal.pbio.0050177. URL http://www.ncbi.nlm.nih.gov/pubmed/17594176. PMID: 17594176.

[81] Poornima Parameswaran, Roxana Jalili, Li Tao, Shadi Shokralla, Baback Gharizadeh, Mostafa Ronaghi, and Andrew Z Fire. A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Research*, 35(19):e130, 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm760. URL http://www.ncbi.nlm.nih.gov/pubmed/17932070. PMID: 17932070.

[82] Matthew R Parsek and E P Greenberg. Sociomicrobiology: the connections between quorum sensing and biofilms. *Trends in Microbiology*, 13(1):27–33, January 2005. ISSN 0966-842X. doi: 10.1016/j.tim.2004.11.007. URL http://www.ncbi.nlm.nih.gov/pubmed/15639629. PMID: 15639629.

[83] Annelie Pernthaler, Anne E Dekas, C Titus Brown, Shana K Goffredi, Tsegereda Embaye, and Victoria J Orphan. Diverse syntrophic partnerships from deep-sea methane vents revealed by direct cell capture and metagenomics. *Proceedings of the National Academy of Sciences of the United States of America*, 105(19):7052–7057, May 2008. ISSN 1091-6490. doi: 10.1073/pnas.0711303105. URL http://www.ncbi.nlm.nih.gov/pubmed/18467493. PMID: 18467493.

[84] Laurent Philippot, Siv G E Andersson, Tom J Battin, James I Prosser, Joshua P Schimel, William B Whitman, and Sara Hallin. The ecological coherence of high bacterial taxonomic ranks. *Nature Reviews. Microbiology*, 8(7):523–529, June 2010. ISSN 1740-1534. doi: 10.1038/nrmicro2367. URL http://www.ncbi.nlm.nih.gov/pubmed/20531276. PMID: 20531276.

[85] Elmar Pruesse, Christian Quast, Katrin Knittel, Bernhard M Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35(21):7188–7196, 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm864. URL http://www.ncbi.nlm.nih.gov/pubmed/17947321. PMID: 17947321.

[86] Dmitry Pushkarev, Norma F Neff, and Stephen R Quake. Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, 27(9):847–852, September 2009. ISSN 1546-1696. doi: 10.1038/nbt.1561. URL http://www.ncbi.nlm.nih.gov/pubmed/19668243. PMID: 19668243.

[87] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H BjÃ¸rn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, SÃ¸ren Brunak, Joel DorÃ©, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, Peer Bork, S Dusko Ehrlich, and Jun Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, March 2010. ISSN 1476-4687. doi: 10.1038/nature08821. URL http://www.ncbi.nlm.nih.gov/pubmed/20203603. PMID: 20203603.

[88] Christopher Quince, Anders Lanzén, Thomas P Curtis, Russell J Davenport, Neil Hall, Ian M Head, L Fiona Read, and William T Sloan. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, 6 (9):639–641, September 2009. ISSN 1548-7105. doi: 10.1038/nmeth.1361. URL http://www.ncbi.nlm.nih.gov/pubmed/19668203. PMID: 19668203.

[89] S Radajewski, P Ineson, N R Parekh, and J C Murrell. Stable-isotope probing as a tool in microbial ecology. *Nature*, 403(6770):646–649, February 2000. ISSN 0028-0836. doi: 10.1038/35001054. URL http://www.ncbi.nlm.nih.gov/pubmed/10688198. PMID: 10688198.

[90] Jeroen Raes and Peer Bork. Molecular eco-systems biology: towards an understanding of community function. *Nature Reviews. Microbiology*, 6(9): 693–699, September 2008. ISSN 1740-1526. doi: 10.1038/nrmicro1935. URL http://www.ncbi.nlm.nih.gov/pubmed/18587409. PMID: 18587409.

[91] Ashna A Raghoebarsing, Arjan Pol, Katinka T van de Pas-Schoonen, Alfons J P Smolders, Katharina F Ettwig, W Irene C Rijpstra, Stefan Schouten, Jaap S Sinninghe Damsté, Huub J M Op den Camp, Mike S M Jetten, and Marc Strous. A microbial consortium couples anaerobic methane oxidation to denitrification. *Nature*, 440(7086):918–921, April 2006. ISSN 1476-4687. doi: 10.1038/nature04617. URL http://www.ncbi.nlm.nih.gov/pubmed/16612380. PMID: 16612380.

[92] Rachna J Ram, Nathan C Verberkmoes, Michael P Thelen, Gene W Tyson, Brett J Baker, Robert C Blake, Manesh Shah, Robert L Hettich, and Jillian F Banfield. Community proteomics of a natural microbial biofilm. *Science (New York, N.Y.)*, 308(5730):1915–1920, June 2005. ISSN 1095-9203. doi: 10.1126/science.1109070. URL http://www.ncbi.nlm.nih.gov/pubmed/15879173. PMID: 15879173.

[93] Sang H Rhee, Charalabos Pothoulakis, and Emeran A Mayer. Principles and clinical implications of the brain-gut-enteric microbiota axis. *Nature Reviews. Gastroenterology & Hepatology*, 6(5):306–314, May 2009. ISSN 1759-5053. doi: 10.1038/nrgastro.2009.35. URL http://www.ncbi.nlm.nih.gov/pubmed/19404271. PMID: 19404271.

[94] M R Rondon, P R August, A D Bettermann, S F Brady, T H Grossman, M R Liles, K A Loiacono, B A Lynch, I A MacNeil, C Minor, C L Tiong, M Gilman, M S Osburne, J Clardy, J Handelsman, and R M Goodman. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied and Environmental Microbiology*, 66(6):2541–2547, June 2000. ISSN 0099-2240. URL http://www.ncbi.nlm.nih.gov/pubmed/10831436. PMID: 10831436.

[95] D J Roser, H J Bavor, and S A McKersie. Application of most-probable-number statistics to direct enumeration of microorganisms. *Applied and Environmental Microbiology*, 53(6):1327–1332, June 1987. ISSN 0099-2240. URL http://www.ncbi.nlm.nih.gov/pubmed/16347364. PMID: 16347364.

[96] J Roth, D LeRoith, M A Lesniak, F de Pablo, L Bassas, and E Collier. Molecules of intercellular communication in vertebrates, invertebrates and microbes: do they share common origins? *Progress in Brain Research*, 68:71–79, 1986. ISSN 0079-6123. URL http://www.ncbi.nlm.nih.gov/pubmed/3562852. PMID: 3562852.

[97] Jonathan M Rothberg and John H Leamon. The development and impact of 454 sequencing. *Nature Biotechnology*, 26(10):1117–1124, October 2008. ISSN 1546-1696. doi: 10.1038/nbt1485. URL http://www.ncbi.nlm.nih.gov/pubmed/18846085. PMID: 18846085.

[98] Edward G Ruby. Symbiotic conversations are revealed under genetic interrogation. *Nature Reviews. Microbiology*, 6(10):752–762, October 2008. ISSN 1740-1534. doi: 10.1038/nrmicro1958. URL http://www.ncbi.nlm.nih.gov/pubmed/18794913. PMID: 18794913.

[99] F Sanger, G M Air, B G Barrell, N L Brown, A R Coulson, C A Fiddes, C A Hutchison, P M Slocombe, and M Smith. Nucleotide sequence of bacteriophage phi x174 DNA. *Nature*, 265(5596):687–695, February 1977. ISSN 0028-0836. URL http://www.ncbi.nlm.nih.gov/pubmed/870828. PMID: 870828.

[100] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, December 1977. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/271968. PMID: 271968.

[101] Patrick D Schloss and Jo Handelsman. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology*, 71(3):1501–1506, March 2005. ISSN 0099-2240. doi: 10.1128/AEM.71.3.1501-1506.2005. URL http://www.ncbi.nlm.nih.gov/pubmed/15746353. PMID: 15746353.

[102] Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, Jason W Sahl, Blaz Stres, Gerhard G Thallinger, David J Van Horn, and Carolyn F Weber. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23): 7537–7541, December 2009. ISSN 1098-5336. doi: 10.1128/AEM.01541-09. URL http://www.ncbi.nlm.nih.gov/pubmed/19801464. PMID: 19801464.

[103] T M Schmidt, E F DeLong, and N R Pace. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *Journal of Bacteriology*, 173(14):4371–4378, July 1991. ISSN 0021-9193. URL http://www.ncbi.nlm.nih.gov/pubmed/2066334. PMID: 2066334.

[104] Sabine P Schrimpf, Manuel Weiss, Lukas Reiter, Christian H Ahrens, Marko Jovanovic, Johan MalmstrÃ¶m, Erich Brunner, Sonali Mohanty, Martin J Lercher, Peter E Hunziker, Ruedi Aebersold, Christian von Mering, and Michael O Hengartner. Comparative functional analysis of the caenorhabditis elegans and drosophila melanogaster proteomes. *PLoS Biology*, 7(3): e48, March 2009. ISSN 1545-7885. doi: 10.1371/journal.pbio.1000048. URL http://www.ncbi.nlm.nih.gov/pubmed/19260763. PMID: 19260763.

[105] H Shizuya, B Birren, U J Kim, V Mancino, T Slepak, Y Tachiiri, and M Simon. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in escherichia coli using an f-factor-based vector. *Proceedings of the National Academy of Sciences of the United States of America*, 89(18):8794–8797, September 1992. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/1528894. PMID: 1528894.

[106] Mitchell L Sogin, Hilary G Morrison, Julie A Huber, David Mark Welch, Susan M Huse, Phillip R Neal, Jesus M Arrieta, and Gerhard J Herndl. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, 103(32):12115–12120, August 2006. ISSN 0027-8424. doi: 10.1073/pnas.0605127103. URL http://www.ncbi.nlm.nih.gov/pubmed/16880384. PMID: 16880384.

[107] D A Stahl, D J Lane, G J Olsen, and N R Pace. Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. *Science (New York, N.Y.)*, 224 (4647):409–411, April 1984. ISSN 0036-8075. doi: 10.1126/science.224.4647.409. URL http://www.ncbi.nlm.nih.gov/pubmed/17741220. PMID: 17741220.

[108] D A Stahl, D J Lane, G J Olsen, and N R Pace. Characterization of a yellowstone hot spring microbial community by 5S rRNA sequences. *Applied and Environmental Microbiology*, 49(6):1379–1384, June 1985. ISSN 0099-2240. URL http://www.ncbi.nlm.nih.gov/pubmed/2409920. PMID: 2409920.

[109] J T Staley and A Konopka. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Review of Microbiology*, 39:321–346, 1985. ISSN 0066-4227. doi: 10.1146/annurev.mi.39.

100185.001541. URL http://www.ncbi.nlm.nih.gov/pubmed/3904603. PMID: 3904603.

[110] Thaddeus S Stappenbeck, Lora V Hooper, and Jeffrey I Gordon. Developmental regulation of intestinal angiogenesis by indigenous microbes via paneth cells. *Proceedings of the National Academy of Sciences of the United States of America*, 99(24):15451–15455, November 2002. ISSN 0027-8424. doi: 10.1073/pnas.202604299. URL http://www.ncbi.nlm.nih.gov/pubmed/12432102. PMID: 12432102.

[111] Bärbel Stecher, Riccardo Robbiani, Alan W Walker, Astrid M Westendorf, Manja Barthel, Marcus Kremer, Samuel Chaffron, Andrew J Macpherson, Jan Buer, Julian Parkhill, Gordon Dougan, Christian von Mering, and Wolf-Dietrich Hardt. Salmonella enterica serovar typhimurium exploits inflammation to compete with the intestinal microbiota. *PLoS Biology*, 5(10):2177–2189, October 2007. ISSN 1545-7885. doi: 10.1371/journal.pbio.0050244. URL http://www.ncbi.nlm.nih.gov/pubmed/17760501. PMID: 17760501.

[112] Bärbel Stecher, Samuel Chaffron, Rina Käppeli, Siegfried Hapfelmeier, Susanne Freedrich, Thomas C Weber, Jorum Kirundi, Mrutyunjay Suar, Kathy D McCoy, Christian von Mering, Andrew J Macpherson, and Wolf-Dietrich Hardt. Like will to like: abundances of closely related species can predict susceptibility to intestinal colonization by pathogenic and commensal bacteria. *PLoS Pathogens*, 6(1):e1000711, January 2010. ISSN 1553-7374. doi: 10.1371/journal.ppat.1000711. URL http://www.ncbi.nlm.nih.gov/pubmed/20062525. PMID: 20062525.

[113] Claire S Ting, Gabrielle Rocap, Jonathan King, and Sallie W Chisholm. Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. *Trends in Microbiology*, 10(3):134–142, March 2002. ISSN 0966-842X. URL http://www.ncbi.nlm.nih.gov/pubmed/11864823. PMID: 11864823.

[114] Susannah G Tringe and Philip Hugenholtz. A renaissance for the pioneering 16S rRNA gene. *Current Opinion in Microbiology*, 11(5):442–446, October 2008. ISSN 1369-5274. doi: 10.1016/j.mib.2008.09.011. URL http://www.ncbi.nlm.nih.gov/pubmed/18817891. PMID: 18817891.

[115] Susannah Green Tringe, Christian von Mering, Arthur Kobayashi, Asaf A Salamov, Kevin Chen, Hwai W Chang, Mircea Podar, Jay M Short, Eric J Mathur, John C Detter, Peer Bork, Philip Hugenholtz, and Edward M Rubin. Comparative metagenomics of microbial communities. *Science (New York, N.Y.)*, 308(5721):554–557, April 2005. ISSN 1095-9203. doi: 10.1126/science.1107851. URL http://www.ncbi.nlm.nih.gov/pubmed/15845853. PMID: 15845853.

[116] Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, Michael Egholm, Bernard Henrissat, Andrew C Heath, Rob Knight, and Jeffrey I Gordon. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, January 2009. ISSN 1476-4687. doi: 10.1038/

nature07540. URL http://www.ncbi.nlm.nih.gov/pubmed/19043404. PMID: 19043404.

[117] Gene W Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E Allen, Rachna J Ram, Paul M Richardson, Victor V Solovyev, Edward M Rubin, Daniel S Rokhsar, and Jillian F Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978): 37–43, March 2004. ISSN 1476-4687. doi: 10.1038/nature02340. URL http://www.ncbi.nlm.nih.gov/pubmed/14961025. PMID: 14961025.

[118] Tim Urich, Anders Lanzén, Ji Qi, Daniel H Huson, Christa Schleper, and Stephan C Schuster. Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PloS One*, 3(6):e2527, 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0002527. URL http://www.ncbi.nlm.nih.gov/pubmed/18575584. PMID: 18575584.

[119] Mark Vellend and Monica A. Geber. Connections between species diversity and genetic diversity. *Ecology Letters*, 8(7):767–781, 2005. doi: 10.1111/j.1461-0248.2005.00775.x. URL http://dx.doi.org/10.1111/j.1461-0248.2005.00775.x.

[120] J Craig Venter, Karin Remington, John F Heidelberg, Aaron L Halpern, Doug Rusch, Jonathan A Eisen, Dongying Wu, Ian Paulsen, Karen E Nelson, William Nelson, Derrick E Fouts, Samuel Levy, Anthony H Knap, Michael W Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jeff Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, and Hamilton O Smith. Environmental genome shotgun sequencing of the sargasso sea. *Science (New York, N.Y.)*, 304(5667):66–74, April 2004. ISSN 1095-9203. doi: 10.1126/science.1093857. URL http://www.ncbi.nlm.nih.gov/pubmed/15001713. PMID: 15001713.

[121] Nathan C VerBerkmoes, Vincent J Denef, Robert L Hettich, and Jillian F Banfield. Systems biology: Functional analysis of natural microbial consortia using community proteomics. *Nature Reviews. Microbiology*, 7(3):196–205, March 2009. ISSN 1740-1534. doi: 10.1038/nrmicro2080. URL http://www.ncbi.nlm.nih.gov/pubmed/19219053. PMID: 19219053.

[122] Nathan C Verberkmoes, Alison L Russell, Manesh Shah, Adam Godzik, Magnus Rosenquist, Jonas Halfvarson, Mark G Lefsrud, Juha Apajalahti, Curt Tysk, Robert L Hettich, and Janet K Jansson. Shotgun metaproteomics of the human distal gut microbiota. *The ISME Journal*, 3(2):179–189, February 2009. ISSN 1751-7370. doi: 10.1038/ismej.2008.108. URL http://www.ncbi.nlm.nih.gov/pubmed/18971961. PMID: 18971961.

[123] C von Mering, P Hugenholtz, J Raes, S G Tringe, T Doerks, L J Jensen, N Ward, and P Bork. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science (New York, N.Y.)*, 315(5815):1126–1130, February 2007. ISSN 1095-9203. doi: 10.1126/science.1133420. URL http://www.ncbi.nlm.nih.gov/pubmed/17272687. PMID: 17272687.

[124] Falk Warnecke and Philip Hugenholtz. Building on basic metagenomics with complementary technologies. *Genome Biology*, 8(12):231, 2007. ISSN 1465-6914. doi: 10.1186/gb-2007-8-12-231. URL http://www.ncbi.nlm.nih.gov/pubmed/18177506. PMID: 18177506.

[125] Christopher M Waters and Bonnie L Bassler. Quorum sensing: cell-to-cell communication in bacteria. *Annual Review of Cell and Developmental Biology*, 21: 319–346, 2005. ISSN 1081-0706. doi: 10.1146/annurev.cellbio.21.012704.131001. URL http://www.ncbi.nlm.nih.gov/pubmed/16212498. PMID: 16212498.

[126] David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, Xavier Gomes, Karrie Tartaro, Faheem Niazi, Cynthia L Turcotte, Gerard P Irzyk, James R Lupski, Craig Chinault, Xing zhi Song, Yue Liu, Ye Yuan, Lynne Nazareth, Xiang Qin, Donna M Muzny, Marcel Margulies, George M Weinstock, Richard A Gibbs, and Jonathan M Rothberg. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189): 872–876, April 2008. ISSN 1476-4687. doi: 10.1038/nature06884. URL http://www.ncbi.nlm.nih.gov/pubmed/18421352. PMID: 18421352.

[127] W B Whitman, D C Coleman, and W J Wiebe. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, 95 (12):6578–6583, June 1998. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/9618454. PMID: 9618454.

[128] Paul Wilmes and Philip L Bond. Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends in Microbiology*, 14(2):92–97, February 2006. ISSN 0966-842X. doi: 10.1016/j.tim.2005.12.006. URL http://www.ncbi.nlm.nih.gov/pubmed/16406790. PMID: 16406790.

[129] C R Woese. Bacterial evolution. *Microbiological Reviews*, 51(2):221–271, June 1987. ISSN 0146-0749. URL http://www.ncbi.nlm.nih.gov/pubmed/2439888. PMID: 2439888.

[130] C R Woese and G E Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, 74(11):5088–5090, November 1977. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/270744. PMID: 270744.

[131] Tanja Woyke, Hanno Teeling, Natalia N Ivanova, Marcel Huntemann, Michael Richter, Frank Oliver Gloeckner, Dario Boffelli, Iain J Anderson, Kerrie W Barry, Harris J Shapiro, Ernest Szeto, Nikos C Kyrpides, Marc Mussmann, Rudolf Amann, Claudia Bergin, Caroline Ruehland, Edward M Rubin, and Nicole Dubilier. Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature*, 443(7114):950–955, October 2006. ISSN 1476-4687. doi: 10.1038/nature05192. URL http://www.ncbi.nlm.nih.gov/pubmed/16980956. PMID: 16980956.

[132] Dongying Wu, Sean C Daugherty, Susan E Van Aken, Grace H Pai, Kisha L Watkins, Hoda Khouri, Luke J Tallon, Jennifer M Zaborsky, Helen E Dunbar,

Phat L Tran, Nancy A Moran, and Jonathan A Eisen. Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *PLoS Biology*, 4(6):e188, June 2006. ISSN 1545-7885. doi: 10.1371/journal.pbio.0040188. URL http://www.ncbi.nlm.nih.gov/pubmed/16729848. PMID: 16729848.

[133] Dongying Wu, Philip Hugenholtz, Konstantinos Mavromatis, RÃ¼diger Pukall, Eileen Dalin, Natalia N Ivanova, Victor Kunin, Lynne Goodwin, Martin Wu, Brian J Tindall, Sean D Hooper, Amrita Pati, Athanasios Lykidis, Stefan Spring, Iain J Anderson, Patrik D'haeseleer, Adam Zemla, Mitchell Singer, Alla Lapidus, Matt Nolan, Alex Copeland, Cliff Han, Feng Chen, Jan-Fang Cheng, Susan Lucas, Cheryl Kerfeld, Elke Lang, Sabine Gronow, Patrick Chain, David Bruce, Edward M Rubin, Nikos C Kyrpides, Hans-Peter Klenk, and Jonathan A Eisen. A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature*, 462(7276):1056–1060, December 2009. ISSN 1476-4687. doi: 10.1038/nature08656. URL http://www.ncbi.nlm.nih.gov/pubmed/20033048. PMID: 20033048.

[134] Jian Xu, Michael A Mahowald, Ruth E Ley, Catherine A Lozupone, Micah Hamady, Eric C Martens, Bernard Henrissat, Pedro M Coutinho, Patrick Minx, Philippe Latreille, Holland Cordum, Andrew Van Brunt, Kyung Kim, Robert S Fulton, Lucinda A Fulton, Sandra W Clifton, Richard K Wilson, Robin D Knight, and Jeffrey I Gordon. Evolution of symbiotic bacteria in the distal human intestine. *PLoS Biology*, 5(7):e156, July 2007. ISSN 1545-7885. doi: 10.1371/journal.pbio.0050156. URL http://www.ncbi.nlm.nih.gov/pubmed/17579514. PMID: 17579514.

[135] Suzan Yilmaz, Mohamed F Haroon, Brian A Rabkin, Gene W Tyson, and Philip Hugenholtz. Fixation-free fluorescence in situ hybridization for targeted enrichment of microbial populations. *The ISME Journal*, May 2010. ISSN 1751-7370. doi: 10.1038/ismej.2010.73. URL http://www.ncbi.nlm.nih.gov/pubmed/20505753. PMID: 20505753.

[136] Kun Zhang, Adam C Martiny, Nikos B Reppas, Kerrie W Barry, Joel Malek, Sallie W Chisholm, and George M Church. Sequencing genomes from single cells by polymerase cloning. *Nature Biotechnology*, 24(6):680–686, June 2006. ISSN 1087-0156. doi: 10.1038/nbt1214. URL http://www.ncbi.nlm.nih.gov/pubmed/16732271. PMID: 16732271.